The Knowledge Engineering Review, Vol. 23:3, 261–281. © 2008, Cambridge University Press
doi:10.1017/S0269888908001379 Printed in the United Kingdom

User-centered evaluation of adaptive and adaptable systems: a literature review

LEX VAN VELSEN, THEA VAN DER GEEST, ROB KLAASSEN and MICHAËL STEEHOUDER

Faculty of Behavioral Sciences, Institute for Behavioral Research, Department of Technical and Professional Communication, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands
e-mail: l.s.vanvelsen@utwente.nl, t.m.vandergeest@utwente.nl, r.klaassen@minaz.nl, m.f.steehouder@utwente.nl

Abstract

This literature review focuses on user-centered evaluation (UCE) studies of adaptive and adaptable systems. Usability, perceived usefulness and appropriateness of adaptation are the three most commonly assessed variables. Questionnaires appeared to be the most popular method, followed by interviews and data log analysis. The quality of most questionnaires was questionable, and the reporting of interviews and think-aloud protocols was found to be shallow. Furthermore, data logs need triangulation in order to be useful. The reports encountered lack empirical value. The article models the iterative design process for adaptive and adaptable systems, linked to the goals of UCE: supporting decisions, detecting problems and verifying quality. This model summarizes the variables to be assessed at each stage of the process and the relevant methods to assess them.

1 Introduction

User-centered evaluation (UCE) can serve three goals: verifying the quality of a product, detecting problems and supporting decisions (De Jong & Schellens, 1997). These functions make UCE a valuable tool for developers of all kinds of systems, because they can justify their efforts, improve upon a system or help developers to decide which version of a system to release. In the end, this may lead to higher adoption of the system, more ease of use and a more pleasant user experience. Figure 1 links the goals of UCE to the phases of the iterative design process. In this article, we discuss the application of UCE to adaptive and adaptable systems on the basis of a literature review and reflect on the UCE practice of such systems.

An *adaptive* system tailors its output, using implicit inferences based on interaction with the user. *Adaptable* systems, in contrast, use explicitly provided input to personalize output. We will use the umbrella term 'personalized systems' to address both types of systems and define them (expanding on the definition of an adaptive system given by Benyon & Murray, 1993) as: systems that can alter aspects of their structure, functionality or interface on the basis of a user model generated from implicit and/or explicit user input, in order to accommodate the differing needs of individuals or groups of users and the changing needs of users over time.

This definition includes the term 'user model'. A user model contains generalizations on a single user or a group of users, based upon, for example, their characteristics, needs or context. The example of book recommendations can illustrate the creation of the user model. Imagine a book shop customer searching for books on sightseeing in Paris, London and Barcelona. The system infers that the customer is interested in city trips, and this characteristic is stored in the user model. Personalized output can then be generated in the form of personal book recommendations, based upon this user model.

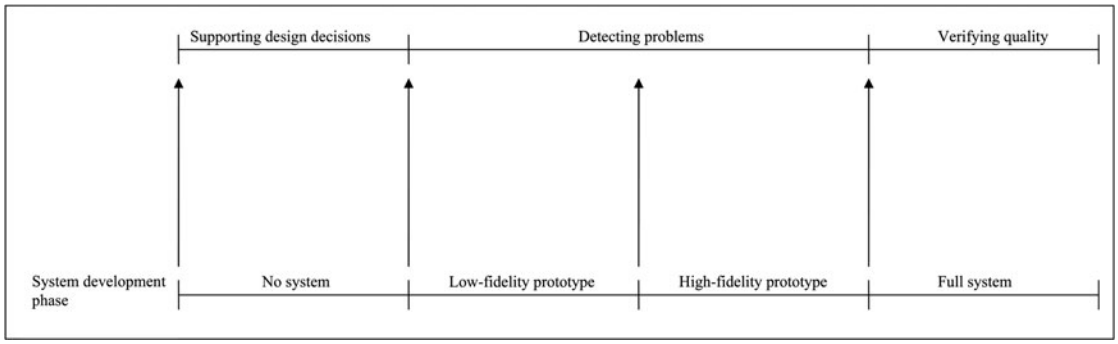

Figure 1 Phases of the iterative design process

'UCE' also needs to be defined. By this we mean an empirical evaluation obtained by assessing user performance and user attitudes toward a system, by gathering subjective user feedback on effectiveness and satisfaction, quality of work, support and training costs or user health and well-being. This definition is partly based on the definition of human-centered design in ISO guideline 13407: 'human-centered design processes for interactive systems' (ISO, 1999).

The inclusion of personalized output causes complications for the UCE of systems. Most traditional evaluation methods are based on the assumption that the system output is the same for each user in every context. When systems include personalized features, this assumption no longer holds. Currently, it is unclear which evaluation methods are best suited for assessing the perceived quality of the unique output of personalized systems. Akoumianakis *et al.* (2001) state that 'The basic qualification criterion, which classifies a technique as being appropriate, is the degree to which it offers the constructs to assess situational and context-specific aspects of user interface adaptations, thus determining its usefulness in a social context' (p. 339). This criterion focuses on the characteristic that is most typical for a personalized system: its adaptation to the user or a group of users and their context. One of the challenges of designing personalized systems is finding a valid, reliable and useful method for UCE. This literature review aims to help address this challenge.

In the past, three surveys on the evaluation of personalized systems have been published. Chin (2001) focuses his overview on the design of experimental evaluations, whereas Gena (2005) and Gena & Weibelzahl (2007) take a more comprehensive view and also include (among others) qualitative evaluation methods in their discussion of empirical evaluation approaches. These surveys have applied a prescriptive approach, describing theory, illustrated with some examples from practice. This survey, however, takes a descriptive approach. After mapping the UCE practice of personalized systems, we will reflect on its quality and provide suggestions for improvement if necessary. A similar approach was used in a concise review by Weibelzahl (2003), where he remarks that the evaluation practice of personalized systems is poor, partly due to non-empirical reporting of activities. However, Weibelzahl does not provide the reader with an elaborate discussion on how malpractice can be avoided or improved upon. This review does seek to inform the practitioner about avoiding the identified pitfalls and the suitability of different methods for the evaluation of personalized systems.

2 Research question and literature selection

The following main question was formulated for the review:

How have user-centered evaluations of personalized systems been conducted in the past and how can this practice be improved upon?

In order to answer this question, the following secondary questions were formulated:

(1) What types of personalized systems have been evaluated?
(2) Which variables have been assessed in the UCE of personalized systems?

(3) Which designs have been used for the UCE of personalized systems?
(4) What kinds of prototypes have been used for the UCE of personalized systems?
(5) Which methods have been used for the UCE of personalized systems?
(6) What are the advantages and disadvantages, the validity and reliability of the methods used for the UCE of personalized systems?
(7) How can the usage of UCE methods when applied to personalized systems be improved upon?

We based our review strategy on the York method (a method for conducting a systematic and empirical literature review), which originates in medical science (NHS Centre for Reviews and Dissemination, 2001). With small modifications (e.g., leaving out descriptions of medical interventions), this method could be applied to the purpose of this study.

We searched for reports on UCE of personalized systems in the following databases or projects: ACM digital library, ERIC, IEEE Xplore, INSPEC, PsycInfo, Science Direct, Scopus, Web of Science, Easy-D (see `http://www.easy-hub.org/hub/studies.jsp` for a list of publications), Peach project (see `http://peach.itc.it/publications.html` for a list of publications). Besides databases, we also consulted the Websites of 15 well-known researchers in the field of personalized systems and searched their publication lists for relevant studies.

The following inclusion criteria were used:

(1) The study had to report the evaluation of a personalized system.
(2) The evaluation of the system had to be (partly) user centered. Studies that only discussed the evaluation of algorithms or other technical aspects of user modeling were excluded.
(3) The study had to describe or discuss at least one of the following issues: advantages or disadvantages, validity, reliability, costs of a method, the participants involved, the variables assessed and the implementation of results in (re)design processes.
(4) The cutoff point was 1990.

For each selected study, the relevant information was assessed and recorded in a database: this included system characteristics (e.g., adaptive feature), research design (e.g., methods used for evaluation) and UCE method characteristics (e.g., validity). The content of this database are available through the Easy-D database on `www.easy-hub.org`.

The search initially resulted in a huge number of hits (>4000). Possibly relevant hits were saved, duplicates were identified and removed, and the abstracts of 338 remaining articles were read. The selection criteria were applied to the abstracts, resulting in 127 studies of which the full text was read. The majority of these studies ($n = 72$) proved not to meet the selection criteria after all, and they were consequently excluded. The 55 remaining studies were included in the review. The Adaptive Hypermedia 2006 conference, which was held shortly after our search, provided six additional studies. Finally, we found two more relevant studies through references in reports (Schmidt-Belz & Poslad, 2003; Cheverst *et al.*, 2005). These were included in our review, making a total of 63 studies. A list of the articles and reports included in the literature review can be found in the appendix.

3 Results

3.1 Systems evaluated in the studies

Most of the studies focused on a single evaluation of an individual system; 19 studies described a series of consecutive evaluations of a system. Only one study reported the evaluation of more than one system. Of the systems included in the review, 23 were adaptive systems (37.1%), 17 were adaptable systems (27.4%) and 22 systems were both adaptive and adaptable (35.5%). In 37 cases, the studies were part of the design process (formative evaluation), whereas 18 studies assessed the appreciation of the system after its implementation (summative evaluation). Eight studies could not be classified as either formative or summative.

Most of the evaluated systems were learning systems (12 studies), followed by intelligent tourist guides (8 studies), information databases (7 studies), interfaces (7 studies) and location-aware

services (7 studies). The PC is the most popular platform for personalized systems, as appeared from the evaluations (42 studies). In 34 studies, the PC application was a Web browser. This does not automatically mean that most personalized systems run via a Website. Two prototype systems were designed as Websites for the sake of evaluation. Website prototypes are relatively easy to develop at low costs. Other platforms we encountered frequently were the PDA (12 studies) and the mobile phone (7 studies).

3.2 User-centered variables assessed in studies

In total, 44 variables were mentioned in the studies. Though different names were used by the authors, the concepts being measured were often identical. Therefore, we grouped these identical concepts and gave them one name. These variables can be grouped in the following categories:

(1) **Variables concerning attitude and experience**

- (a) Appreciation
- (b) Trust and privacy issues
- (c) User experience
- (d) User satisfaction

(2) **Variables concerning actual use**

- (a) Usability
- (b) User behavior
- (c) User performance

(3) **Variables concerning system adoption**

- (a) Intention to use
- (b) Perceived usefulness

(4) **Variables concerning system output**

- (a) Appropriateness of adaptation
- (b) Comprehensibility
- (c) Unobtrusiveness

In the studies in our review, 'usability' was most frequently measured, followed by 'perceived usefulness' and 'appropriateness of adaptation'. Table 1 shows how often each variable was addressed in the 63 reviewed studies. The wordings of most variables speak for themselves. We will however, define usability: 'the extent to which a product can be used by specified users to

Table 1 Variables addressed in UCE in collected studies ($n = 63$)

Variable		Times included in evaluation
1	Usability	33
2	Perceived usefulness	26
3	Appropriateness of adaptation	26
4	Intention to use	25
5	User behavior	24
6	Appreciation	18
7	User satisfaction	18
8	User performance	17
9	Comprehensibility	10
10	User experience	8
11	Trust and privacy issues	7
12	Unobtrusiveness	1
13	Other variables	4

achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use' (BSI, 1998).

4 User-centered evaluation designs and their application to personalized systems

This section deals with two common UCE designs: comparisons and laboratory or real-life settings.

4.1 Comparisons: personalized versus non-personalized systems

Comparisons try to identify the differences between a personalized version of a system and one without the personalizing feature.

4.1.1 Design usage

Fourteen studies compared a personalized and a non-personalized version of a system. Most of these (8 out of 14) measured the *user performance* with two versions of a system (e.g., the amount of learning done with a system). The goal was to judge whether the personalized system was better for the user than the non-personalized system.

4.1.2 Design implications

The validity and reliability of such comparisons has been discussed at length in the literature (e.g., Höök, 1997, 2000). Comparing a personalized system with one where the personalization has been removed is deemed a false comparison. When the personalized features have been removed, the system is no longer a worthy opponent (Höök, 2000). The evaluation reports we found also discussed this issue. Furthermore, users may be biased toward the personalized version of the system, they may favor new technology, or they may want to please the experimenter and give socially desirable answers (Bohnenberger *et al*., 2002). Finally, the study designs appeared to consider too short an interaction time to understand the full effect of personalization on the user. In order to identify an effect, longitudinal studies are needed (Höök, 1997).

4.1.3 Comment

There are serious questions what a comparison shows. If, for example, a personalized system is found to be more efficient than a non-personalized version, this may suggest that the personalized version is better. But the comparison does not tell much about usability, perceived output, or future adoption. The practical value of comparisons with respect to user-centered design is limited (Alpert & Vergo, 2007). Furthermore, defining quality in terms of user performance is troublesome. If it takes users longer to accomplish a task with a non-personalized version of a system, but they have a better experience using that version, one has to wonder which version is 'better'. Is there an absolute 'better' version? Evaluators must be aware of the limited importance of efficiency when it comes to the total user experience.

Comparing personalized and non-personalized systems can be used to determine user performance in comparison with another system. However, it needs to be very clear *what* is being compared with *what*. If a non-personalized system is involved, this system should be a worthy equivalent for the personalized system and not a weak version of the original system. For example, comparing a personalized system and information provided on paper (see Cawsey *et al*., 2000) may be fairer, and it can provide valuable insights as well.

4.2 Laboratory and real-life observations

When observing, a researcher watches a participant working with a personalized system, noting interesting events, or recording the whole session.

4.2.1 Design usage

In our collection of studies, laboratories were mentioned seven times, but none of the reports specifies how the laboratory was used. It is therefore difficult to determine whether the use of a

laboratory was a plus or not. Five studies used real-life observations, which can assess the system in the realm of 'real world issues'.

4.2.2 Design implications

Using a laboratory allows the researcher to control the environment. For empirical research, this permits one to exclude outside influences, so one can focus on the variables one wants to assess. The downside of using a laboratory is that one loses the real-life setting; using the laboratory reduces ecological validity. Because the situation in the laboratory is not identical to a real-life setting, results may not be generalizable. According to the basic criteria for the appropriateness of a UCE method for personalized systems (Akoumianakis *et al.*, 2001; see Chapter 1), laboratory observations are not ideal, because the (social) context in which interaction normally takes place may be difficult to simulate in a laboratory.

4.2.3 Comment

For some systems (e.g., a personalized Website), a laboratory may provide a valid and reliable evaluation. The choice of an artificial or real-life testing environment depends on the relation the system has with its surroundings. If real-life settings are important to the working of a system, it might be wise to test the system in the field instead of the laboratory.

5 Prototypes used for user-centered evaluation of personalized systems

In order to evaluate a system, subjects need to interact with a (prototype) version of the system. Otherwise, subjects have difficulty imagining how the system will work and how it will relate to their everyday activities (Weibelzahl *et al.*, 2006). The types of prototypes used in the reported studies are displayed in Table 2. The first three will be discussed in detail. Since not all reports specified what system was used for evaluations, it is difficult to assess whether a prototype was used or not. For a complete overview of the different prototypes one can use during system design, refer to Maguire (2001).

The use of a working system prototype was reported 17 times. Creating a working prototype is relatively easy and cheap compared with creating a full system, and it can help to verify the quality of a product. Creating a prototype is a wise use of money and effort (Field *et al.*, 2001). However, the use of a prototype version of a system does not yield the same results as one would obtain using the full system (Field *et al.*, 2001). A simplified version may therefore not be the right instrument for identifying usability issues.

A computer simulation offers the participant the possibility to interact with it (e.g., in Gena & Torre, 2004). Simulations are a feasible means of testing a system when development is well under way, but not yet completed. Nevertheless, one has to be cautious when generalizing the results and applying them to another device. People may interact differently with a computer than with other devices (Buchauer *et al.*, 1999; Goren-Bar *et al.*, 2005).

Low-fidelity prototypes allow the researcher to test personalization in a very early phase of system development (e.g., in Karat *et al.*, 2003). This can pinpoint crucial issues that may play a role in the adoption of the system in the future and can help to assess the participants' attitudes (Walker *et al.*, 2002). Low-fidelity prototypes often include just a few screenshots of the system,

Table 2 Prototypes used for UCE of personalized systems

Interaction method	Times used	Percentage of total ($n = 63$) (%)
Working prototype	17	27.0
Computer simulation	3	4.8
Paper prototype	3	4.8
Mockup simulation	2	3.2
Wizard-of-Oz prototype	2	3.2
Demonstration of the system	2	3.2

Table 3 Methods used for UCE in the collected studies ($n = 63$)

Evaluation method or instrument	Times applied	Variables most frequently assessed
Questionnaires*	47	Usability (21) Perceived usefulness (19) Intention to use (14)
Interviews*	27	Intention to use (9) Appropriateness of adaptation (7) Usability (7)
Data log analysis	24	User behavior (17) User performance (6)
Focus groups* & Group discussions*	8	Intention to use (4) Perceived usefulness (2) Trust & privacy issues (2)
Think-aloud protocols*	7	Usability (2) User behavior (2)
Expert reviews	6	Usability (3)

Only methods that are used more than twice are shown.
*Purely UCE methods.

which are not necessarily representative of the final version. They must provide the participant with an idea of the system functions and (personalized) output (Virzi *et al*., 1996). Participants can comment on the concept behind the system or discuss the functions the system comprises. The abstract nature of paper prototypes may produce abstract user feedback. Therefore, in low-fidelity prototyping, one needs methods that are flexible and can cover aspects in depth, like interviews and focus groups.

6 User-centered evaluation methods and their application to personalized systems

The methods and instruments used in the investigated studies are listed in Table 3. Two points are worth noting. First, almost every study uses multiple evaluation methods. Second, some of the listed methods (e.g., data log analysis) may not seem to be user centered according to our definition of UCE, because they do not collect subjective feedback from (potential) users. Studies that used methods that are not user centered along with methods that are user centered were still included in our review. In these cases, the methods should be seen as complementary, providing triangulated data, and contributing to the overall user-centeredness of the evaluation.

In the following subsections the methods listed in Table 3 will be discussed more thoroughly.

6.1 Questionnaires

A questionnaire collects data from respondents by letting them answer a fixed set of questions, either on paper or on screen. Items on these questionnaires can be closed (when participants can choose one of multiple choices) or open (when participants can freely answer in writing).

6.1.1 Method usage

Questionnaires were the most common evaluation method in the studies, and were used 47 times. The most commonly measured variables were *usability* (21 times), *perceived usefulness* (19 times) and *intention to use* (14 times). Interestingly, questionnaires were often used both for gathering global impressions of the system and as a tool to identify problem areas. In 25 studies, the questionnaire was used as a form of formative evaluation and in 16 studies for summative evaluation.

Table 4 Respondents per questionnaire

Number of respondents	Encountered in literature review
0–25	15 times
25–100	23 times
100–250	6 times
>250	4 times

One of the advantages of questionnaires is the large number of participants that can be accommodated (compared with interviews or other methods). Table 4 shows the different sample sizes. The use of questionnaires completed by only a small number of subjects is questionable from a methodological perspective. If a sample group is small, it is difficult to generalize findings (Weibelzahl *et al.*, 2002). A small number of respondents limits the statistical processing, and possible presumed relations, which could have been significant with a larger sample size, may turn out to be insignificant. Some authors discussed the need for a large group of respondents in order to generate a meaningful evaluation (Cawsey *et al.*, 2000) or commented that their study suffered from the small number of respondents (Henderson *et al.*, 1998; Gregor *et al.*, 2003). In order to make claims about the usability of a system, for example, sample sizes need to be large enough that they can be generalized to a larger population of users (Dicks, 2002). In the evaluation design, the designated number of participants should be geared to the goals of the evaluation. Whereas problem detection can be done with a relatively small sample, the verification of quality needs a larger, representative group of respondents.

Another issue worth mentioning is the construction of the questionnaires themselves. Before items can be constructed, the variable one wants to investigate needs to be defined. Next, the items measuring this variable have to be geared upon this definition. An example of mismatch between variable and items can be found in Cheverst *et al.* (2005). In their evaluation of an intelligent control system, they measured the variable 'Acceptance' by means of items on 'previous experience with such systems, (e.g., automatic doors, Microsoft Office Assistant) in different environments (i.e., home, work)' (p. 262). The authors do not define 'Acceptance' whereas the items do not assess the essence of the variable as it is used in the field of technology adoption and where acceptance comprises high usage of a technology by designated users. The items by Cheverst *et al.* assess 'experience with similar systems'. Although this variable might influence acceptance, there is certainly more to acceptance than previous experience alone. The article however, does not clarify whether the variable is defined wrong and should be assessed by means of different items, or should be named 'Previous experience with similar systems' because the definition of the variable is in compliance with the items.

Measuring psychological constructs with only one or two items (e.g., in Henderson *et al.*, 1998) is generally not considered good practice, and measuring two concepts in one question might easily yield non-valid results (Spector, 1992). An example of the latter is when Stary & Totter (2003) asked participants: 'Overall, how clear and useful are the contents of the information system?' Clarity and usefulness are two distinct variables, but here the participants must express them in one answer on a Likert scale. As a result, no definite answer can be expected. Finally, the design of a Likert scale should result in a measurement continuum (Spector, 1992). Muntean & McManis (2006), for example, designed a five point Likert scale to assess usability and user satisfaction. Their scale included the response choices: 1, poor; 2, fair; 3, average; 4, good and 5, excellent. The inclusion of 'fair' as the number 2 choice distorts the continuum of the Likert scale. Therefore, one cannot interpret an average score on one of these items as a valid indication of usability or user satisfaction: the results are skewed toward a neutral or positive reaction.

The results of questionnaires are not reported systematically and as a result, it may be hard for a reader to judge the quality of an evaluation. When assessing quantitative scores on variables by means of multiple items and Likert scales, it is good practice to first determine and report the

quality of measurement. By determining and reporting Cronbach's α (Cronbach, 1951) for each variable, readers can assess the quality of a questionnaire. Next, reporting the means and standard deviations for each variable can provide the reader with a clear view on the answers the participants gave and their consensus (or lack thereof) on a certain topic.

6.1.2 Comment

Questionnaires may not be the best choice for identifying problems of usability. A qualitative evaluation method, like thinking aloud, will be more suitable for this purpose (Van Velsen *et al.*, 2007). However, a questionnaire can be very useful for measuring appreciation of personalization, user satisfaction, general opinions about the system, or for benchmarking purposes. Such variables are suitable for quantitative measurement, using Likert scales, for example.

Effective questionnaires can yield useful insights, hopefully leading to the construction of a standard quantitative instrument, based on questionnaires used in the past. Therefore, we encourage the publication of the used questionnaires in appendices or on Websites. Furthermore, we encourage evaluators to design and validate questionnaires according to the guidelines listed by Spector (1992) and report them according to the guidelines of Kitchenham *et al.* (2002).

6.2 Interviews

In interviews, participants are asked questions in person by an interviewer. Interviews can be structured, with fixed questions, or semi-structured, if the interviewer can also ask questions that come up during the interview.

6.2.1 Method usage

Twenty-seven studies reported interviews with users. This makes it the second most frequently used method. In 19 studies, the interview could be qualified as part of the formative evaluation, and in 7 as part of the summative evaluation. In interviews, the variable *intention to use* was addressed nine times, and *appropriateness of adaptation* and *usability* were each addressed seven times. No other variables stand out.

The manner in which interview results have been reported, make it seem that evaluators consider interviews to be inferior to statistical data. Gregor *et al.* (2003), for example, conducted a study by means of effectiveness tests and interviews. In their results section, they only discuss the effectiveness of the system. The interview results are only marginally mentioned in the discussion section, where they say 'Comments from the dyslexic pupils indicated a strong subjective preference for their individually selected settings over the defaults of the word processor' (p. 353). Because the authors do not point out how many participants gave this indication (e.g., three out of six), we have to take their word for this conclusion. Only reporting trends in interview results is a phenomenon we have encountered often (e.g., Sodergard *et al.*, 1999; Ketamo, 2003; Gena & Torre, 2004; Kolari *et al.*, 2004). When analyzing interviews, one can apply several systematic approaches, for example, analyzing results per question asked, or steps in a process (e.g., teachers' perceptions of steps involved in learning students how to engineer requirements). In order to prevent confusion, it is best to choose one approach and stick to it (Patton, 2002). Results from interviews deserve a systematic analysis and presentation in the proper section of the report: the results section. They are too valuable to be treated as side issue.

In almost all cases, interviews were conducted after the interviewee had used the system. As a result, experiences and possibly important comments might be forgotten or not deemed relevant, and thus never be reported. A solution for this problem might be an interview with the help of video as used by Goren-Bar *et al.* (2005). They videotaped the subject while interacting with a personalized museum guide and interviewed the subject afterward, while watching the recording together. The rationale behind this method was that the researchers hoped to discover the reasoning behind user actions. A positive side-effect may be that users remember and mention experiences they might otherwise have forgotten.

6.2.2 Comment

The interview is a feasible method for assessing system adoption (Van Velsen *et al.*, 2007). Intention to use and perceived usefulness are mostly the product of individual and situational factors which interviews can identify; this can help explain why a system will or will not be adopted. If more were known about these factors, it might be possible to develop a theory that explains these two phenomena, or to adjust theories like the Technology Acceptance Model (Davis *et al.*, 1989) or the Unified Theory of Acceptance and Use of Technology (Venkatesh *et al.*, 2003) for personalized systems. Interviews may also be used to assess the usability issues that are typical for personalized systems since it allows detailed feedback. As a result, not only can problem areas be identified, but also perceived causes and solutions.

The reports of interviews in the articles were generally not very detailed. The interviewer and his background often remain unmentioned and systematic overviews of the data collection process and results are scarce. In reports, the processing of data from the interviews has to be described and justified, in order to determine its quality. For an overview on how to analyze qualitative data, like interview results, we refer to Miles & Huberman (1994). When reporting interviews, a good standard is to provide typical comments made by interviewees (e.g., Ketamo, 2003; Smith *et al.*, 2004).

6.3 Data log analysis

Data logs record the actions performed by participants when interacting with a system. These data logs can serve as the basis for quantitative or qualitative analysis. The analysis can focus on user behavior (clickstream analysis) or the user performance.

6.3.1 Method usage

The interaction with the system was recorded in 24 studies. Data logs were most frequently used to assess user behavior (17 times), and in some cases user performance (6 times).

One strength of the method lies in the possibility of collecting huge amounts of data without disturbing the user (the method is unobtrusive). Another advantage is that data logging records behavior objectively.

6.3.2 Practical implications

Analyzing data logs to determine user performance may be troublesome. Stein (1997) reports that it may be difficult to interpret task completion time as a measure of user performance (the faster a subject completes a task, the more efficient a system), as personalized output may sometimes provide the user with more (difficult) output on purpose (e.g., when the system assumes a student needs more reading material to understand a lesson). Furthermore, data logs often cover only a relatively small period of interaction time with the system. As a consequence, no information is gathered about the growing competence users acquire while working with a system (Höök, 1997).

One of the most important drawbacks of using data logs as an evaluation method is that this does not provide insight into the causes of problems that are discovered (Jensen *et al.*, 2000). For example, one can conclude from a set of logs that some users skip the introduction page of a system and immediately try to execute tasks. As a result, they may need more time than users who did not skip the introduction page. Data logs do not provide information on *what made* the user skip the introduction page. In order to answer these kinds of questions and to gain a full understanding of user behavior, data log analysis should be combined with other (qualitative) evaluation methods, like thinking aloud or focus groups. Triangulation can help to determine the reasons behind the user behavior, which is necessary to generate successful improvements (Herder, 2006).

Finally, there are ethical questions associated with collecting data logs. Internet browsing is considered a personal activity and logging user behavior may be regarded as an infringement of privacy (Herder, 2006).

6.3.3 Comment

We strongly advise using data log analysis to assess user behavior or user performance in combination with a qualitative and UCE method. The use of triangulation provides deeper understanding of results generated with data log analysis (Ammenwerth *et al.*, 2003). A possible setup of such a study would be to make people work with the system first and to collect and analyze data logs. Then, peculiarities or questions that have arisen during data log analyses can be presented or posed to focus group participants. This way, effects *and* their causes can be established which is valuable information for future system (re)design.

6.4 Focus groups and group discussions

In focus groups and group discussions, a group of participants discusses a fixed set of topics or questions. These discussions are led by a moderator, who can ask questions that come up during the session.

6.4.1 Method usage

Eight studies used focus groups or group discussions, mostly focused on the intention to use a system (four times).

The reporting of focus group results lacks the same detail as we discussed in Section 6.2. Evaluators do not systematically report the answers provided by the participants, but present trends. Hyldegaard & Seiden (2004), for example, state that 'site information, actuality, language and layout of the interface were often used by the participants as arguments for trusting or not trusting a personalized output'. From such a statement, we cannot assess whether any of these arguments were mentioned more often and therefore, are perhaps more important.

Furthermore, the setup of focus group is reported very summarily and as a result, the empirical value of a study suffers. We illustrate this with the setup of Chesnais *et al.* (1995) which goes: 'We performed focus studies during Spring 1994 to gauge the acceptance of the system' (p. 280). When readers do not understand what has been done, they cannot repeat the study or learn from the design for eventual future use. Because of such limited explanations, we could not judge the quality of these sessions well.

6.4.2 Practical implications

A typical issue with focus groups is that the moderator almost entirely influences the nature of the discussion and the discussed topics. As a result, participants' answers are only in line with these topics. Comments on other topics, which may be of great importance, are seldom expressed by participants (Morgan, 1996).

6.4.3 Comment

Focus groups and group discussions are suitable for gathering a large amount of qualitative data in a short time. They may help to provide a deeper understanding of user behavior, appreciation, and intention to use, as we illustrated in Section 6.3 with our example on the combination of data logging and focus groups. Another interesting and promising approach is to combine focus groups and paper prototyping. This can generate useful results in an early phase of the development process (Kaasinen, 2003; Karat *et al.*, 2003).

6.5 Think-aloud protocols

When thinking aloud, participants are asked to use a system and say their thoughts out loud.

6.5.1 Method usage

Seven studies used think-aloud protocols. The quality of reporting of think-aloud protocols was poor. Researchers often assume that mentioning the use of such protocols is enough for the reader to understand what is being done. It would be a positive addition to evaluation reports if writers would mention which variables were assessed by means of think-aloud protocols. Moreover, thinking aloud often happens in congruence with test tasks. These test tasks determine the scope

of the user–system interaction during the evaluation and therefore, can influence results. In order to guarantee a valid evaluation that takes into account all the main aspects of the system, the test tasks need to address them. When presenting the evaluation methodology, test tasks need to be clarified to improve the empirical value of the study and to show a well-balanced distribution of system functions over test tasks. Statements like 'Visitors were asked to perform eight tasks in total. They were also asked to think aloud while interacting with the system' (Pateli *et al.*, 2005 p. 203) do not suffice.

6.5.2 Practical implications

Thinking aloud may increase a participant's cognitive load and, as a result, influence task performance (Van den Haak *et al.*, 2003). This increased cognitive load can make it difficult for participants to verbalize non-task related problems, because they have to give their full attention to interacting with a system (Van den Haak *et al.*, 2004). In this case, observations of the participant can make up for the loss of verbalizations.

6.5.3 Comment

Think-aloud protocols can be fruitful instruments for understanding user behavior, especially the user's rationale behind it. It might also be a useful way to generate feedback on usability issues (Van Velsen *et al.*, 2007). How to setup a thinking-aloud session and to analyze the resulting protocol can be found in Benbunan-Fich (2001) and Ericsson & Simon (1993).

6.6 Expert reviews

In an expert review, an expert studies a system and gives his or her view on it.

6.6.1 Method usage

Expert reviews were present in six studies, mostly to establish usability. In some reports, the setup of the expert reviews and their results are discussed to a satisfactory level (e.g., Gates *et al.*, 1998), while in some reports their procedure remains vague or even unknown (e.g., Kaasinen, 2003).

6.6.2 Practical implications

One can doubt whether expert reviews are 'user centered'. Literature indicates that experts are not capable of supplying the same critique on a system that (potential) end-users can. Van der Geest (2004) found that expert reviews uncover different issues than focus groups or think-aloud protocols (both of which involve end-users). Lentz & De Jong (1997) showed that experts are unable to predict user problems because experts have different skills, knowledge, cultural backgrounds and interests than users. Hartson *et al.* (2003) agree, stating that the realness of usability problems identified by experts can only be determined by real users. Finally, a study by Savage (1996) identified a discrepancy between expert and user reviews. Although experts identified areas of an interface that needed to be further investigated, users focused their comments on changes that had to be made to the interface.

The 'evaluator effect' determines the usability issues uncovered by an expert review. The varied skills and experience of experts affect the results of the evaluation (Hertzum & Jacobsen, 2003; Hvannberg *et al.*, 2007). It is assumed that the more skilled or experienced an expert is, the better the evaluation. This effect can be eliminated by using a large group of experts.

The application of heuristics is often implied in expert reviews. Heuristics are design rules for a system, and a heuristic evaluation determines whether a system has been designed in accordance with these rules. Applying heuristics is closely connected to expert reviews. Expert reviews and heuristic evaluation are in essence both applications of a set of design rules (either implicitly stored in an expert's memory or explicitly written down) to a system. Kjeldskov *et al.* (2005) showed that applying textbook heuristics is too static to assess the usability of personalized systems. It is an inappropriate method, as it is unable to account for different user characteristics and contexts, and thus uncovers only a limited number of usability problems. Experts may 'suffer'

from the same limitation, being unable to identify with every user and with every context the personalized system is designed for. However, Magoulas *et al*. (2003) showed that heuristics, altered to fit the system, can be useful when formulating redesign guidelines.

6.6.3 Comment

There are drawbacks in using expert reviews or heuristics to evaluate personalization. Evaluations conducted with (potential) end-users will probably yield more valuable results, because they can account for the different characteristics and contexts involved. On the other hand, expert review or heuristics may be valuable tools for evaluating specific topics not involving personalization, like accessibility and legal issues.

7 Discussion

7.1 User-centered evaluation in the iterative design process

On the basis of the discussion above, we model the iterative design process for a personalized system in Figure 2. A similar overview per method has been published by Gena & Weibelzahl (2007).

We distinguish four phases, each one focusing on a more mature (prototype) version of the system. We linked these phases to the goals UCE can serve, following De Jong & Schellens (1997). In each phase, a set of variables can be assessed, depending on the version of the system that participants are interacting with. The literature review has shown which methods are most useful for assessing the different variables. Phases, variables and methods are linked in Figure 2. This indicates what is most appropriately investigated in each phase and the most appropriate method for this.

The model could help evaluators and researchers to choose the right variables and methods for their evaluation. By gearing the variables and methods to the goals of a study, meaningful evaluations can be conducted (e.g., problems with usability can be detected by means of think-aloud protocols). The model can also serve as a framework through which one can view evaluation studies or the evaluation process.

In the first phase, there is no system or prototype at hand, so studies can only assess the context in which the system is supposed to function, or gather features the end-user would like to see in it; nothing can be determined yet about how users appreciate the system (or a version of it). Therefore, 'requirements engineering' is a better term than evaluation at this stage. As this is a whole discipline in itself, discussing it lies outside the scope of this review. However, requirements engineering is a substantial part of the iterative design process of personalized systems, and contributes significantly to the system design.

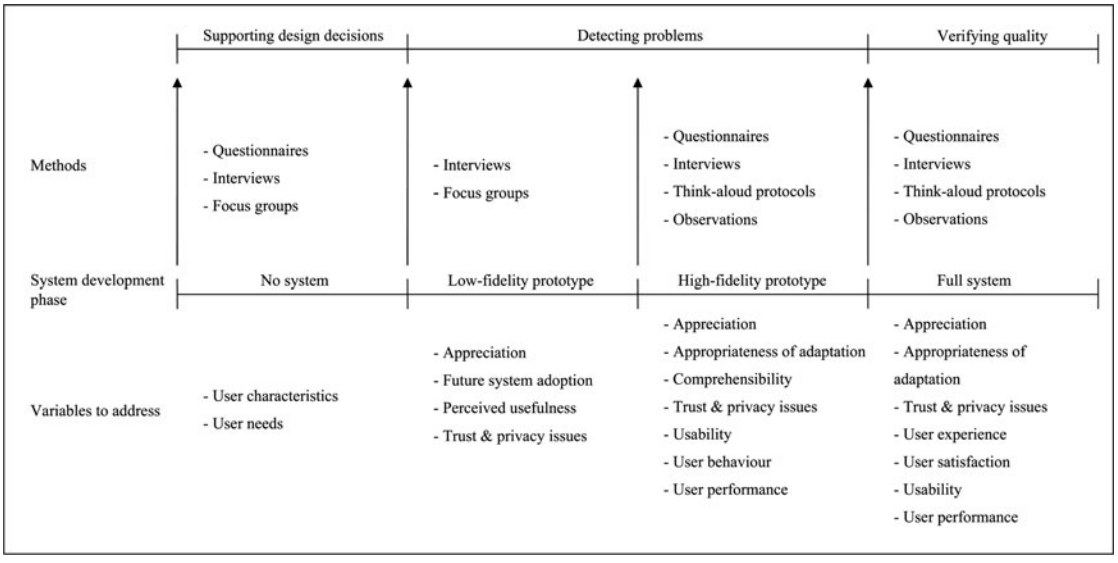

Figure 2 The iterative design process and the role of user-centered evaluation for personalized systems

The second phase in the development process deals with the ideas behind the system and the techniques that are supposed to embody these ideas. A low-fidelity prototype can visualize them, and from this moment on, UCE becomes an option. The methods that are most appropriate here are those that can collect in-depth data. Qualitative methods, like interviews or focus groups, can be used to assess perceived usefulness, future system adoption, appreciation and trust in the system being developed. Trust in a system is an important topic for personalized systems, and should be addressed at an early phase in the development process (Kramer *et al.*, 2000).

In the third phase, a working prototype can support evaluations that address both usability and (adapted) output. Although this phase is characterized by a 'working prototype', the system may already be in a well-advanced stage. Quantitative, as well as qualitative evaluation methods may be beneficial now, depending on the variables one wants to assess.

The fourth phase involves testing the finished system in a real-life environment. The transition from formative (prototype) to summative (full system) evaluation means a change in evaluation goals. Where a prototype evaluation focuses on problem-oriented results, a full system evaluation will focus on benchmark results. User satisfaction, user experience and trust in the system are examples of variables that can be assessed at this phase, for example, with a questionnaire.

7.2 Implications for future research and evaluations

This review investigated which methods are suitable for UCE of personalized systems. At this moment, the pros and cons of different methods of evaluating personalization have not been fully documented. As mentioned in Chapter 1, Akoumianakis *et al.* (2001) state that 'The basic qualification criterion, which classifies a technique as being appropriate, is the degree to which it offers the constructs to assess situational and context-specific aspects of user interface adaptations, thus determining its usefulness in a social context' (p. 339). This is a good starting point for exploring the applicability of different methods. It seems clear that more research is needed concerning UCE methods and their application to personalized systems. It would be interesting to compare two research methods evaluating the same system. However, researching and comparing methods has limitations, because each personalized system functions in a unique environment, which means that we must be cautious about generalizing the results. The distinct users and different contexts to which the system must adapt may have such a large influence that it may be necessary to assess a method's validity, reliability, advantages and disadvantages for every appearance of the system, and also perhaps for different contexts.

When we reflect on the methods which have been applied during evaluations, we see that questionnaires are extremely popular. This approach is radically different from those normally used in user-centered design practices (Vredenburg *et al.*, 2002). Some of the appended questionnaires we encountered in our review were poorly designed, which might be cause for concern. Questionnaires seem to be used as the quick and dirty way of assessing user opinion about personalized systems. As a result, the potential of questionnaires is not fully exploited, and the quality of the resulting evaluations is low. The reason for using a questionnaire needs to be considered carefully. Issues that call for an extensive review of the system, like usability, require other, more exhaustive methods. Questionnaires with closed items can only confirm known variables and assess their scores. Other methods, such as interviews or focus groups, are more exploratory, starting with a clean slate and letting the respondents' answers guide the results. Because little is known about the interaction between user and personalized systems, it may be best to use exploratory methods to assess important variables. In a later phase, these variables can serve as input for confirming methods (like questionnaires).

The empirical value of many evaluation reports was very limited. Evaluation designs were often not specified well enough to make it possible for evaluators to replicate the study. This makes it also hard to judge the quality of an evaluation (Weibelzahl, 2005). If we do not know what happened, we cannot take the results, or the conclusions, for granted. In order to improve upon the empirical value of a study and a report, we advice evaluators to use the guidelines for research and reporting as discussed by Kitchenham *et al.* (2002).

One reason for the low quality of some evaluations may be that most evaluators of personalized systems are computer scientists and not specialized in evaluation (Weibelzahl, 2005). In order to increase the quality of evaluations and their results, system developers will have to be educated in evaluation or they will have to hire specialists to assist them during the setup of the study and the analysis of results. The latter option also prevents a bias from being included in the evaluation, as system developers may find it difficult to speak negatively about a product they developed themselves.

A topic that has not been discussed in depth in this article is the measurement of effectiveness of personalized systems. An overview of the metrics one should apply for the evaluation of a personalized system would be valuable and useful for the community. However, such metrics differ per domain, as each domain has a specific set of goals that needs to be achieved. Within a domain, these goals can differ even further from system to system. As the measurement of success is dependent on this goal, it is difficult, if not impossible, to present the definite set of 'metrics for successful personalization'. From a user-centered point of view, we prefer to determine the success of a personalized system by means of the goal it is supposed to serve, as perceived by the user. A similar view has previously been advocated by Weibelzahl (2005). We want to illustrate it by means of the example of a personalized search engine. In this domain, popular system-centered metrics are 'precision' and 'recall' (Su, 1992). Precision deals with the percentage of returned results on a query that is relevant, and recall with the percentage of all relevant results that are presented to the user after submitting a query. Whether a result is relevant is judged by experts. However, is a personalized search engine successful if it achieves better scores on precision and recall than a non-personalized version? What a user may want from a search engine is to receive relevant results, and to receive them fast. Therefore, the metric of success should be measured by means of perceived relevance of search results, as suggested by Nahl (1998), combined with the time or mouse clicks it takes searchers to get to these results. Then, to determine the success of personalization, scores can be determined for a personalized version of the search engine and a non-personalized variant. As we mentioned in Section 4.1, this competing system must be a fair opponent. Therefore, in this case, the fair opponent might best be a popular, non-personalized search engine like *Google* or *Yahoo!*.

There are some variables whose importance we would like to stress and which can be assessed for the majority of personalized systems, namely usability issues for personalized systems. The reports we read did not generally assess these usability issues (predictability, controllability, unobtrusiveness, privacy and breadth of experience), as compiled by Jameson (2003, 2006). Because these issues determine the shape of the user experience, they can prevent the user from wanting to interact with a personalized system (e.g., if the user perceives the collection of user data as a serious infringement on privacy). Assessing usability issues is an important part of a UCE of a personalized system.

8 Summary and conclusion

Using a literature review, we have commented on the use of UCE methods for the evaluation of personalized systems. The current UCE practice of personalized systems was found to be sloppy at times. Based on previous approaches, we identified different variables and some effective methods of assessing them. Questionnaires may be suitable for assessing general statements or for benchmarking purposes. We advocate that questionnaire items be published with the study or on Websites. Interviews may be useful to obtain information on the intention to use a system or to determine its usability. The design and analysis of interviews should be reported in more detail. Data log analysis can inform us about user behavior, but it should be used in combination with a qualitative method to determine the rationale behind the behavior. Focus groups can help to establish the rationale behind user behavior or the intention to use. Think-aloud protocols can provide feedback on user behavior as well as usability. As with interviews, researchers should report how they employed think-aloud protocols in more detail. Finally, expert reviews or heuristics can assess issues outside personalization, like accessibility or legal issues. However, one should be cautious about using expert reviews or heuristics as a substitute for user testing.

UCE can be of great value in the iterative design process of personalized systems because it improves the interaction between user and system. At this moment, we do not know a great deal about the suitability of different evaluation methods, due to the relatively early stage of research into personalized systems. This review gives evaluators insights into how to apply methods and which variables to assess at different phases in the system design process. Furthermore, this review and the database, published on EASY-D, can provide a starting point for more reflection on UCE practices. However, it is crucial that evaluators evade well-known pitfalls and that writers of future evaluation reports increase their empirical value, by reporting the used methodology and results in such a fashion that replication of the study is possible. We are convinced that the quality of evaluations will benefit and that, indirectly, the user will be served in the process.

References

Akoumianakis, D., Grammenos, D. & Stephanidis, C. 2001. User interface adaptation: Evaluation perspectives. In *User interfaces for All*, Stephanidis, C. (ed.). Erlbaum, 339–352.

Alpert, S. R. & Vergo, J. G. 2007. User-centered evaluation of personalized web sites: What's unique? In *Human Computer Interaction Research in Web Design and Evaluation*, Zaphiris, P. & Kurniawan, S. (eds). Idea Group, 257–272.

Ammenwerth, E., Iller, C. & Mansmann, U. 2003. Can evaluation studies benefit from triangulation? A case study. *International Journal of Medical Informatics* **70**(2/3), 237–248.

Benbunan-Fich, R. 2001. Using protocol analysis to evaluate the usability of a commercial web site. *Information & Management* **39**(2), 151–163.

Benyon, D. & Murray, D. 1993. Adaptive systems: From intelligent tutoring to autonomous agents. *Knowledge-Based Systems* **6**(4), 197–219.

Bohnenberger, T., Jameson, A., Krüger, A. & Butz, A. 2002. Location-aware shopping assistance: Evaluation of a decision-theoretic approach. *Lecture Notes in Computer Science* **2411**, 155–169.

BSI. 1998. *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals*. British Standards Institution.

Buchauer, A., Pohl, U., Kurzel, N. & Haux, R. 1999. Mobilizing a health professional's Workstation: Results of an evaluation study. *International Journal of Medical Informatics* **54**, 105–114.

Cawsey, A. J., Jones, R. B. & Pearson, J. 2000. The evaluation of a personalised health information system for patients with cancer. *User Modeling and User-Adapted Interaction* **10**, 47–72.

Chesnais, P. R., Mucklo, M. J. & Sheena, J. A. 1995. The Fishwrap personalized news system. *Paper Presented at the Second International Workshop on Community Networking 'Integrated Multimedia Services to the Home'*, Princeton, USA.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C. & Villar, N. 2005. Exploring issues of user model transparency and proactive behaviour in an office environment control system. *User Modeling and User-Adapted Interaction* **15**, 235–273.

Chin, D. N. 2001. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction* **11**, 181–194.

Cronbach, L. J. 1951. Coefficient alpha and the internal structure of tests. *Psychometrika* **16**(3), 297–334.

Davis, F. D., Bagozzi, R. P. & Warshaw, P. R. 1989. User acceptance of computer technology: A comparison of two theoretical models. *Management Science* **35**, 982–1003.

De Jong, M. D. T. & Schellens, P. J. 1997. Reader-focused text evaluation. An overview of goals and methods. *Journal of Business and Technical Communication* **11**(4), 402–432.

Dicks, R. S. 2002. Mis-usability: On the uses and misuses of usability testing. *Paper Presented at the 20th Annual International Conference on Computer Documentation*, Toronto, Canada.

Ericsson, K. A. & Simon, H. A. 1993. *Protocol Analysis*. MIT Press.

Field, A., Hartel, P. & Mooij, W. 2001. Personal DJ, an open architecture for personalised content delivery. *Paper Presented at the Tenth International Conference World Wide Web*, Hong Kong.

Gates, K. F., Lawhead, P. B. & Wilkins, D. E. 1998. Toward an adaptive www: A case study in customised hypermedia. *New Review of Hypermedia and Multimedia* **4**, 89–113.

Gena, C. & Torre, I. 2004. The importance of adaptivity to provide onboard services: A preliminary evaluation of an adaptive tourist information service onboard vehicles. *Applied Artificial Intelligence* **18**(6), 549–580.

Gena, C. 2005. Methods and techniques for the evaluation of user-adaptive systems. *Knowledge Engineering Review* **20**(1), 1–37.

Gena, C. & Weibelzahl, S. 2007. Usability engineering for the adaptive web. In *The Adaptive Web*, Brusilovsky, P., Kobsa, A. & Nejdl, W. (eds). Springer-Verlag, 720–762.

Goren-Bar, D., Graziola, I., Kuflik, T., Pianesi, F., Rocchi, C., Stock, O. & Zancanaro, M. 2005. I like it: An affective interface for a multimodal museum guide. Retrieved on 20 April 2006 from `http://peach.itc.it/papers/gorenbar2005.pdf`

Gregor, P., Dickinson, A., Macaffer, A. & Andreasen, P. 2003. Seeword: A personal word processing environment for dyslexic computer users. *British Journal of Educational Technology* **34**(3), 341–355.

Hartson, H. R., Andre, T. S. & Williges, R. C. 2003. Criteria for evaluating usability evaluation methods. *International Journal of Human-Computer Interaction* **15**(1), 145–181.

Henderson, R., Rickwood, D. & Roberts, P. 1998. The beta test of an electronic supermarket. *Interacting with Computers* **10**, 385–399.

Herder, E. 2006. *Forward, Back and Home Again. Analyzing User Behaviour on the Web*. University of Twente.

Hertzum, M. & Jacobsen, N. E. 2003. The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction* **15**(1), 183–204.

Höök, K. 1997. Evaluating the utility and usability of an adaptive hypermedia system. *Paper Presented at IUI '97*, Orlando, USA

Höök, K. 2000. Steps to take before intelligent user interfaces become real. *Interacting with Computers* **12**(4), 409–426.

Hvannberg, E. T., Law, E. L., Lárusdóttir, M. K. 2007. Heuristic evaluation: Comparing ways of finding and reporting usability problems. *Interacting with Computers* **19**, 225–240.

Hyldegaard, J. & Seiden, P. 2004. My E-journal: Exploring the usefulness of personalized access to scholarly articles and services. *Information Research* **9**(3), paper 181. Available at `http://InformationR.net/ir/9-3/paper181.html`

ISO. 1999. *ISO 13407: Human-Centered Design Processes for Interactive Systems*, International Standard Organizations.

Jameson, A. 2003. Adaptive interfaces and agents. *In Human-Computer Interaction Handbook*, Jacko, J. A. & Sears A. (eds). Erlbaum, 305–330.

Jameson, A. 2006. Adaptive interfaces and agents. In *Human-Computer Interaction Handbook* (2nd edn), Jacko, J. A. & Sears, A. (eds). Erlbaum, 433–458.

Jensen, A. L., Boll, P. S., Thysen, I. & Pathak, B. K. 2000. Pl@nteInfo®: A web-based system for personalised decision support in crop management. *Computers and Electronics in Agriculture* **25**, 271–293.

Kaasinen, E. 2003. User needs for location aware mobile services. *Personal Ubiquitous Computing* **7**(1), 70–79.

Karat, C. M., Brodie, C., Karat, J., Vergo, J. & Alpert, S. R. 2003. Personalizing the user experience on ibm. Com. *IBM Systems Journal* **42**(4), 686–701.

Ketamo, H. 2003. Xtask: An adaptable learning environment. *Journal of Computer Assisted Learning* **19**(3), 360–370.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K. & Rosenberg, J. 2002. Preliminary guidelines for empirical research in software engineering. *IEEE Transactions on Software Engineering* **28**(8), 721–734.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S. & Davies, J. 2005. Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behaviour & Information Technology* **24**(1), 51–65.

Kolari, J., Laakko, T., Hiltunen, T., Ikonen, V., Kulju, M., Suihkonen, R., Toivonen, S., & Virtanen, T. 2004. *Context-Aware Services for Mobile Users (Technology and User Experiences)*, VTT publications 539.

Kramer, J., Noronha, S. & Vergo, J. 2000. A user-centered design approach to personalization. *Communications of the ACM* **43**(8), 45–48.

Lentz, L. & De Jong, M. D. T. 1997. The evaluation of text quality: Expert-focused and reader-focused methods compared. *IEEE Transactions on Professional Communication* **40**(3), 224–234.

Magoulas, G. D., Chen, S. Y. & Papanikolaou, K. A. 2003. Integrating layered and heuristic evaluation for adaptive learning environments. *Paper Presented at the Second Workshop on Empirical Evaluation of Adaptive Systems Held in Conjunction with User Modelling 2003*, Pittsburg, USA.

Maguire, M. 2001. Methods to support human-centred design. *International Journal of Human-Computer Studies* **55**(4), 587–634.

Miles, M. B. & Huberman, A. M. 1994. *Qualitative Data Analysis* (2nd edn), Sage.

Morgan, D. L. 1996. Focus groups. *Annual Review of Sociology*, **22**, 129–152.

Muntean, C. H. & McManis, J. 2006. The value of QoE-based adaptation approach in educational hypermedia: Empirical evaluation. *Lecture Notes in Computer Science* **4018**, 121–130.

Nahl, D. 1998. Ethnography of novices' first use of web search engines: Affective control in cognitive processing. *Internet Reference Services Quarterly* **3**(2), 51–72.

NHS Centre for Reviews and Dissemination. 2001. *Undertaking Systematic Reviews of Research on Effectiveness*, University of York.

Pateli, A. G., Giaglis, G. M. & Spinellis, D. D. 2005. Trial evaluation of wireless info-communication and indoor location-based services in exhibition shows. *Advances in Informatics* **3746**, 199–210.

Patton, M. Q. 2002. Qualitative Research & Evaluation Methods, (3rd edn), Sage.

Savage, P. 1996. User interface evaluation in an iterative design process: A comparison of three techniques. *Paper presented at CHI'96*, Vancouver, Canada.

Schmidt-Belz, B. & Poslad, S. 2003. User validation of a mobile tourism service. *Paper Presented at the Workshop on HCI in Mobile Guides*, Udine, Italy.

Smith, H., Fitzpatrick, G. & Rogers, Y. 2004. Eliciting reactive and reflective feedback for a social communication tool: A multi-session approach. *Paper Presented at Designing Interactive Systems: Across the Spectrum*, Cambridge, USA.

Sodergard, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Jarvinen, T., Kaasinen, E., Kinnunen, T., Kolari, J., Kunnas, J. & Tammela, A. 1999. Integrated multimedia publishing: Combining TV and newspaper content on personal channels. *Computer Networks: the International Journal of Computer and Telecommunications Networking* **31**, 1111–1128.

Spector, P. E. 1992. *Summated Rating Scale Construction*, Sage.

Stary, C. & Totter, A. 2003. Measuring the adaptability of universal accessible systems. *Behaviour & Information Technology* **22**(2), 101–116.

Stein, A. 1997. Usability and assessments of multimodal interaction in the SPEAK! System: An experimental case study. *The New Review of Hypermedia and Multimedia* **3**, 159–180.

Su, L. T. 1992. Evaluation measures for interactive information retrieval. *Information Processing and Management* **28**(4), 503–516.

Van den Haak, M. J., De Jong, M. D. T. & Schellens, P. J. 2003. Retrospective vs. concurrent think-aloud protocols: Testing the usability of an online library catalogue. *Behaviour & Information Technology* **22**(5), 339–351.

Van den Haak, M. J., De Jong, M. D. T. & Schellens, P. J. 2004. Employing think-aloud protocols and constructive interaction to test the usability of online library catalogues: A methodological comparison. *Interacting with Computers* **16**, 1153–1170.

Van der Geest, T. 2004. Beyond accessibility: Comparing three website usability test methods for people with impairments. *Paper Presented at HCI. 2004: Design for Life*, Leeds, England.

Van Velsen, L., Van der Geest, T. & Klaassen, R. 2007. Testing the usability of a personalized system: Comparing the use of interviews, questionnaires and thinking-aloud. *Paper Presented at the IEEE Professional Communication Conference*, Seattle, USA.

Venkatesh, V., Morris, M. G., Davis, G. B. & Davis, F. D. 2003. User acceptance of information technology: Towards a unified view. *MIS Quarterly* **27**(3), 425–478.

Virzi, R. A., Sokolov, J. L. & Karis, D. 1996. Usability problem identification using both low- and high-fidelity prototypes. *Paper Presented at the SIGHI Conference on Human Factors in Computing Systems: Common Ground*, Vancouver, Canada.

Vredenburg, K., Mao, J., Smith, P. W. & Carey, T. 2002. A survey of user-centered design practice. *CHI Letters* **4**(1), 471–478.

Walker, M., Takayama, L. & Landay, J. A. 2002. High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes. *Paper Presented at the 46th Annual Meeting of the Human Factors and Ergonomics Society*, Baltimore, USA.

Weibelzahl, S., Lippitsch, S. & Weber, G. 2002. Advantages, opportunities, and limits of empirical evaluations: Evaluating adaptive systems. *Künstliche Intelligenz* **3**(2), 17–20.

Weibelzahl, S. 2003. *Evaluation of Adaptive Systems. PhD thesis*, University of Trier.

Weibelzahl, S. 2005. Problems and pitfalls in the evaluation of adaptive systems. In *Adaptable and Adaptive Hypermedia Systems*, Chen, S. Y. & Magoulas, G. D. (eds). IRM Press, 285–299.

Weibelzahl, S., Jedlitschka, A. & Ayari, B. 2006. Eliciting requirements for a adaptive decision support system through structured user interviews. *Paper Presented at the User Centered Design and Evaluation Workshop Held in Conjunction with Adaptive Hypermedia '06*, Dublin, Ireland.

Appendix: Reports included in the review

Alpert, S. R., Karat, J., Karat, C. M., Brodie, C. & Vergo, J. G. 2003. User attitudes regarding a user-adaptive eCommerce web site. *User Modeling and User-Adapted Interaction* **13**, 373–396.

Bange, M. P., Deutscher, S. A., Larsen, D., Linsley, D. & Whiteside, S. 2004. A handheld decision support system to facilitate improved insect pest management in Australian cotton systems. *Computers and Electronics in Agriculture* **43**(2), 131–147.

Baumgartner, P., Furbach, U., Gross-Hardt, M. & Sinner, A. 2004. Living book: Deduction, slicing and interaction. *Journal of Automated Reasoning* **32**(3), 259–286.

Biemans, M. C. M., Van Kranenburg, H. & Lankhorst, M. 2001. User evaluations to guide the design of an extended personal service environment for mobile services. *Paper Presented at the 5th international symposium on wearable computers*, Zurich, Switzerland.

Bloom, C., Linton, F.. & Bell, B. 1997. Using evaluation in the design of an intelligent tutoring system. *Journal of Interactive Learning Research* **8**(2), 235–276.

Bohnenberger, T., Jameson, A., Krüger, A. & Butz, A. 2002. Location-aware shopping assistance: Evaluation of a decision-theoretic approach. *Lecture Notes in Computer Science* **2411**, 155–169, Springer, Herdelberg.

Brusilovsky, P. & Anderson, J. 1998. Act-r electronic bookshelf: An adaptive system to support learning act-r on the web. *Paper Presented at Webnet 98 World Conference of the WWW, Internet, and Intranet,* Orlando, USA.

Cawsey, A. J., Jones, R. B. & Pearson, J. 2000. The evaluation of a personalised health information system for patients with cancer. *User Modelling and User-Adapted Interaction* **10**, 47–72.

Chesnais, P. R., Mucklo, M. J. & Sheena, J. A. 1995. The Fishwrap personalized news system. *Paper Presented at the Second International Workshop on Community Networking 'Integrated Multimedia Services to the Home',* Princeton, USA.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C. & VIllar, N. 2005. Exploring issues of user model transparency and proactive behaviour in an office environment control system. *User Modelling and User-Adapted Interaction* **15**, 235–273.

Cheverst, K., Davies, N., Mitchell, K., Friday, A., Efstratiou, C. 2000. Developing a context-aware electronic tourist guide: Some issues and experiences. *CHI Letters* **2**(1), 17–24.

Conlan, O., O'Keeffe, I. & Tallon, S. 2006. Combining adaptive hypermedia techniques and ontology reasoning to produce dynamic personalized news services. *Lecture Notes in Computer Science* **4018**, 81–90, Springer, Herdelberg.

Conlan, O. & Wade, V. P. 2004. Evaluation of APeLS: An adaptive elearning service based on the multi-model, metadata-driven approach. *Lecture Notes in Computer Science* **3137**, 291–295, Springer, Herdelberg.

Cox, R., O'Donnel, M. & Oberlander, J. 1999. Dynamic versus static hypermedia in museum education: An evaluation of ilex, the intelligent labelling explorer. *Paper Presented at the Artificial Intelligence in Education Conference*, Le Mans, France.

De Almeida, P. & Yokoi, S. 2003. Interactive character as a virtual tour guide to an online museum exhibition. In *Museums and the Web 2003: Selected Papers from an International Conference*, Bearman, D. & Trant, J. (eds). Archives & Museum Informatics, Pittsbrugh, 191–198.

De Roure, D., Hall, W., Reich, S., Hill, G., Stairmand, M. & Pikrakis, A. 2001. Memoir: An open framework for enhanced navigation of distributed information. *Information Processing and Management* **37**(1), 53–74.

Díaz, A., Gervás, P. & García, A. 2005. Evaluation of a system for personalized summarization of web contents. *Lecture Notes in Computer Science* **3538**, 453–462, Springer, Herdelberg.

Díaz, A., Gervas, P., García, A. & Chacon, I. 2001. Sections, categories and keywords as interest specification tools for personalised news services. *Online Information Review* **25**(3), 149–159.

Eason, K., Yu, L. & Harker, S. 2000. The use and usefulness of functions in electronic journals: The experience of the SuperJournal project. *Program* **34**, 1–28.

Field, A., Hartel, P. & Mooij, W. 2001. Personal DJ, an open architecture for personalised content delivery. *Paper Presented at the Tenth International Conference World Wide Web*, Hong Kong.

Gates, K. F., Lawhead, P. B. & Wilkins, D. E. 1998. Toward an adaptive www: A case study in customised hypermedia. *New Review of Hypermedia and Multimedia* **4**, 89–113.

Gena, C. 2002. An empirical evaluation of an adaptive web site. *Paper Presented at IUI'02*, San Francisco, USA.

Gena, C. & Torre, I. 2004. The importance of adaptivity to provide onboard services: A preliminary evaluation of an adaptive tourist information service onboard vehicles. *Applied Artificial Intelligence* **18**(6), 549–580.

Goren-Bar, D., Graziola, I., Kuflik, T., Pianesi, F., Rocchi, C., Stock, O. & Zancanaro, M. 2005. I like it: An affective interface for a multimodal museum guide. Retrieved on 20 April 2006 from `http://peach.itc.it/papers/gorenbar2005.pdf`

Goren-Bar, D. & Kuflik, T. 2004. Don't miss-r: Recommending restaurants through an adaptive mobile system. *Paper Presented at the Ninth International Conference on Intelligent User Interface*, Funchal, Portugal.

Graziola, I., Pianesi, F., Zancanaro, M. & Goren-Bar, D. 2005. Dimensions of adaptivity in mobile systems: Personality and people's attitudes. *Paper Presented at IUI'05*, San Diego, USA.

Gregor, P., Dickinson, A., Macaffer, A. & Andreasen, P. 2003. Seeword: a personal word processing environment for dyslexic computer users. *British Journal of Educational Technology* **34**(3), 341–355.

Hatala, M. & Wakkary, R. 2005. Ontology-based user modeling in an augmented audio reality system for museums. *User Modeling and User-adapted Interaction* **15**(3/4), 339–380.

Henderson, R., Rickwood, D. & Roberts, P. 1998. The beta test of an electronic supermarket. *Interacting with Computers* **10**, 385–399.

Höök, K. 1997. Evaluating the utility and usability of an adaptive hypermedia system. *Paper Presented at IUI '97*, Orlando, USA.

Hyldegaard, J. & Seiden, P. 2004. My E-journal: exploring the usefulness of personalized access to scholarly articles and services. *Information Research* **9**(3), paper 181. Available at `http://InformationR.net/ir/9-3/paper181.html`

Isobe, T., Fujiwara, M., Kaneta, H., Morita, T. & Uratani, N. 2005. Development of a TV reception navigation system personalized with viewing habits. *IEEE Transactions on Consumer Electronics* **51**(2), 665–674.

Isobe, T., Fujiwara, M., Kaneta, H., Uratani, N. & Morita, T. 2003. Development and features of a TV navigation system. *IEEE Transactions on Consumer Electronics* **49**(4), 1035–1042.

Järvinen, T. (ed.) 2005. *Hybridmedia as a Tool to Deliver Personalised Product-Specific Information About Food. Report of the TIVIK Project*, VTT publications, 2304.

Kaasinen, E. 2003. User needs for location aware mobile services. *Personal Ubiquitous Computing* **7**(1), 70–79.

Karat, C. M., Brodie, C., Karat, J., Vergo, J. & Alpert, S. R. 2003. Personalizing the user experience on ibm. Com. *IBM Systems Journal* **42**(4), 686–701.

Kavcic, A., Privosnik, M., Marolt, M. & Divjak, S. 2002. Educational hypermedia: An evaluation study. *Paper Presented at the Proceedings of the Mediterranean Electrotechnical Conference—MELECON*, Cairo, Egypt.

Ketamo, H. 2003. Xtask: An adaptable learning environment. *Journal of Computer Assisted Learning* **19**(3), 360–370.

Kiris, E. 2004. User-centered eservice design and redesign. *Paper Presented at CHI 2004*, Vienna, Austria.

Kolari, J., Laakko, T., Hiltunen, T., Ikonen, V., Kulju, M., Suihkonen, R., Toivonen, S., & Virtanen, T. 2004. *Context-Aware Services for Mobile Users Technology and User Experiences*, VTT publications, 539.

Krauss, F. S. H. 2003. Methodology for remote usability activities: A case study. *IBM Systems Journal* **42**(4), 582–593.

Kuiper, P. M., Van Dijk, E. M. A. G. & Boerma, A. K. 2006. Adaptive municipal e-forms. *Paper Presented at the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, Held in Conjunction with the 4th International Conference on Adaptive Hypermedia & Adaptive Web Based Systems (AH'06)*, Dublin, Ireland.

Masthoff, J. 2006. The user as wizard: A method for early involvement in the design and evaluation of adaptive systems. *Paper Presented at the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, Held in Conjunction with the 4th International Conference on Adaptive Hypermedia & Adaptive Web Based Systems (AH'06)*, Dublin, Ireland.

McGrenere, J., Baecker, R. M. & Booth, K. S. 2002. An evaluation of a multiple interface design solution for bloated software. *CHI Letters* **4(1)**, 163–170.

Muntean, C. H. & McManis, J. 2006. The value of QoE-based adaptation approach in educational hypermedia: Empirical evaluation. *Lecture Notes in Computer Science* **4018**, 121–130, Springer, Herdelberg.

O'Grady, M. J. & O'Hare, G. M. P. 2004. Enabling customized & personalized interfaces in mobile computing. *Paper Presented at IUI'04*, Madeira, Portugal.

Patel, D. & Marsden, G. 2004. Customizing digital libraries for small screen devices. *Proceedings of the 2004 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries*, Stellenbosch, Western Cape, South Africa: South African Institute for Computer Scientists and Information Technologists.

Pateli, A. G., Giaglis, G. M. & Spinellis, D. D. 2005. Trial evaluation of wireless info-communication and indoor location-based services in exhibition shows. *Advances in Informatics* **3746**, 199–210.

Phillips, M., Hawkins, R., Lunsford, J. & Sinclair-Pearson, A. 2004. Online student induction: A case study of the use of mass customization techniques. *Open Learning* **19**(2), 191–202.

Romero, C., Ventura, S., Hervas, C., & De Bra, P. 2006. An authoring tool for building both mobile adaptable tests and web-based adaptive or classic tests. *Lecture Notes in Computer Science* **4018**, 203–212, Springer, Herdelberg.

Schmidt-Belz, B. & Poslad, S. 2003. User validation of a mobile tourism service. *Paper Presented at the Workshop on HCI in Mobile Guides*, Udine, Italy.

Smith, H., Fitzpatrick, G. & Rogers, Y. 2004. Eliciting reactive and reflective feedback for a social communication tool: A multi-session approach. *Paper Presented at Designing Interactive Systems: Across the Spectrum*, Cambridge, USA.

Smyth, B. & Cotter, P. 2000. A personalized television listings service. *Communications of the ACM* **43**(8), 107–111.

Sodergard, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Jarvinen, T., Kaasinen, E., Kinnunen, T., Kolari, J., Kunnas, J. & Tammela, A. 1999. Integrated multimedia publishing: Combining TV and newspaper content on personal channels. *Computer Networks: the International Journal of Computer and Telecommunications Networking* **31**, 1111–1128.

Stary, C. & Totter, A. 2003. Measuring the adaptability of universal accessible systems. *Behaviour & Information technology* **22**(2), 101–116.

Stathis, K., De Bruijn, O. & Macedo, S. 2002. Living memory: Agent-based information management for connected local communities. *Interacting with Computers* **14**(6), 663–688.

Stein, A. 1997. Usability and assessments of multimodal interaction in the SPEAK! System: An experimental case study. *The New Review of Hypermedia and Multimedia* **3**, 159–180.

Storey, M. A., Phillips, B., Maczewski, M. & Wang, M. 2002. Evaluating the usability of web-based learning tools. *Educational Technology and Society* **5**(3), 91–100.

Thomas, C. G. 1993. Design, implementation and evaluation of an adaptive user interface. *Knowledge-Based Systems* **6**(4), 230–238.

Virvou, M. & Alepis, E. 2005. *Mobile educational features in authoring tools for personalised tutoring*. Computers & Education **44**(1), 53–68.

Weibelzahl, S. 2003. *Evaluation of Adaptive Systems*. Pedagogical University Freiburg.

Weibelzahl, S., Jedlitschka, A. & Ayari, B. 2006. Eliciting requirements for a adaptive decision support system through structured user interviews. *Paper Presented at the Fifth Workshop on User-Centered Design and Evaluation of Adaptive Systems, Held in Conjunction with the 4th International Conference on Adaptive Hypermedia & Adaptive Web Based Systems (Atl '06)*, Dublin, Ireland.

Yue, W., Mu, S., Wang, H. & Wang, G. 2005. TGH: A case study of designing natural interaction for mobile guide systems. *Proceedings of the 7th International Conference on Human Computer Interaction with Mobile Devices & Services*, Salzburg, Austria.