

Usability Engineering for the Adaptive Web

Cristina Gena¹ and Stephan Weibelzahl²

¹ Dipartimento di Informatica, Università di Torino
Corso Svizzera 185, Torino, Italy
`cgena@di.unito.it`

² School of Informatics, National College of Ireland
Mayor Street, Dublin, Ireland
`sweibelzahl@ncirl.ie`

Summary. This chapter discusses a usability engineering approach for the design and the evaluation of adaptive web-based systems, focusing on practical issues. A list of methods will be presented, considering a user-centered approach. After having introduced the peculiarities that characterize the evaluation of adaptive web-based systems, the chapter describes the evaluation methodologies following the temporal phases of evaluation, according to a user-centered approach. Three phases are distinguished: requirement phase, preliminary evaluation phase, and final evaluation phase. Moreover, every technique is classified according to a set of parameters that highlight the practical exploitation of that technique. For every phase, the appropriate techniques are described by giving practical examples of their application in the adaptive web. A number of issues that arise when evaluating an adaptive system are described, and potential solutions and workarounds are sketched.

1.1 Introduction

Involving users in the design and evaluation of adaptive web-based systems has the potential to considerably improve the systems' effectiveness, efficiency and usability. Many authors have emphasized the importance of empirical studies [32, 64, 65, 96, 154], as well as the lack of suitable examples reported in the literature. Like most systems, adaptive web-based systems [100] can benefit considerably from user involvement in design and evaluation.

1.1.1 The User's Perspective for System Design

Designing adaptive web-based systems is challenging from a usability perspective [65, 73], because some of the inherent principles of these systems (e.g., automatically tailoring the interface) might violate standard usability principles such as user control and consistency (see Section 1.4.5).

Usability engineering is the systematic process of developing user interfaces that are easy to use [109, 159]. A variety of methods have been proposed to

ensure that the interface of the final product is efficient to use, easy to learn, and satisfying to use. This includes heuristics and guidelines, expert reviews, and user-centered design methods. The rationale of user-centered design (UCD) is to place the user as opposed to the software artifact, at the center of the design process [77]. Users are involved in the development process in very early phases of the software development and in fact throughout the complete development life-cycle. Involving users from the very beginning can help to discover their ideas and expectations about the system (the so-called mental model). Moreover, it can help to identify and analyze tasks, workflows and goals, and in general to validate the developers' assumptions about the users.

As usability engineering and user centered design methods focus on cognitive and ergonomic factors (such as perception, memory, learning, problem-solving, etc.) they seem particularly suitable for the design of user-adaptive systems. The anticipation and the prevention of usability side effects should form an essential part of the iterative design of user-adaptive systems [74]. Many of these methods are described throughout this chapter. Before applying them though, we strongly encourage readers to still consult a "practical" textbook on user needs analysis and evaluation, such as [36].

1.1.2 The User's Perspective for System Evaluation

Evaluation is the systematic determination of merit, worth, and significance of something. In software development, evaluations are used to determine the quality and feasibility of preliminary products such as mock-ups and prototypes as well as of the final system. It also has the advantage of providing useful feedback for subsequent redesigns.

Adaptive systems adapt their behavior to the user and/or the user's context. The construction of a user model usually requires making many assumptions about users' skills, knowledge, needs or preferences, as well as about their behavior and interaction with the system. Empirical evaluation offers a way of testing these assumptions in the real world or under more controlled conditions [154]. Evaluation results can offer valuable insights about the real behavior and preferences of users. They can demonstrate that a certain adaptation technique actually works, i.e., that it is accurate, effective, and efficient. Evaluation studies are an important means to convince users, customers or investors of the usefulness and feasibility of a system. Finally, evaluations are important for scientific advancement as they offer a way to compare different approaches and techniques.

1.1.3 Formative versus Summative Evaluation

Often evaluation is seen as the final mandatory stage of a project. While the focus of many project proposals is on new theoretical considerations or some innovative features of an adaptive system, a summative evaluation study is often planned in the end as empirical validation of the results. However, when

constructing a new adaptive system, the whole development cycle should be covered by various evaluation studies, from the gathering of requirements to the testing of the system under development (see Sec. 1.3.2).

Formative evaluations are aimed at checking the first design choices before actual implementation and getting the clues for revising the design in an iterative design-re-design process.

From this perspective, evaluation can be considered as a *generative method* [43], since it offers contributions during the design phase by providing the means of combining design specification and evaluation into the same framework. Evaluation results can offer insights about the real behavior and the preferences of users, and therefore be adopted in the construction of the user models and system adaptation mechanisms. Expert and real users are a strong source of information for the knowledge base of the system and their real behavior offers insight for the intelligent behavior of the system. Therefore, as will be demonstrated, in adaptive web systems, evaluation is important not only to test usability and functionality, but also because testing methodologies can be a knowledge source for the development of the adaptivity components (e.g., user data acquisition, interface adaptations, inference mechanisms, etc).

The focus of this chapter is on practical issues for carrying out adaptive web-based system evaluation under a usability engineering point of view, suggesting methods and criteria to help researchers and students that are faced with evaluation problems. Since evaluation is still a challenge, we have to promote appropriate testing methodologies and publish empirical results that can be generalized, in order to check the effectiveness of adaptive web systems and put them into practice.

The chapter presents a comprehensive overview of empirical and non-empirical methods focusing on the peculiarities of the web. A detailed list of techniques will be presented, derived from Human Computer Interaction (HCI). In the evaluation of adaptive systems, especially in the final evaluation phase, metrics from information retrieval and information filtering systems are used (e.g., accuracy of recommendations, accuracy of system predictions and/or system preferences, similarity of expert rating and system prediction, inferred domain knowledge in the user model, etc) in order to evaluate the effectiveness of content adaptations. As far as these methodologies are concerned, relevant surveys are already available [19, 28, 55, 132], and Chapter 12 of this book shows examples of recommender systems evaluation [29].

therefore the main focus of the chapter will be on those HCI methods which are used in the iterative design-evaluation process. These are often disregarded in the adaptive web, even if they can contribute to an improvement in the evaluation of adaptive web systems.

When designing empirical studies on adaptive web-based systems a number of typical issues may arise. Section 1.4 provides an overview of these issues and suggests possible solutions or workarounds.

1.2 The Proposed Approach to the Analysis of Evaluation Techniques

In order to produce effective results, evaluation should occur throughout the entire design life cycle and provide feedback for design modifications [109, 159]. Early focus on users and tasks, continual testing of different solution-prototypes, empirical measurement, and integrated and iterative design can help to avoid expensive design mistakes. All the mentioned principles are also the key-factors of the user-centered design approach [114]: to involve users from the first design decisions of an interactive system and to understand the user's needs and address them in very specific ways. Gould and Lewis [56] originally phrased this principle as follows:

- early focus on users and tasks;
- empirical measurements of product usage;
- iterative design in the production process.

A more direct engagement with final users can help to discover the context in which interaction takes place. This is particularly important both when considering ethnographic approaches (see Sec. 1.3.3) and when investigating the adaptation to the context in adaptive web sites for portable devices.

Since we believe that the usability engineering methodologies and the user-centered approach can become key factors for successful design and evaluation of adaptive web systems, in this chapter the evaluation techniques will be listed according to the life-cycle stage in which they can occur: requirement phase, preliminary evaluation phase, and final evaluation phase.

1.2.1 Classification of Techniques

In order to give some practical suggestions, at the end of every technique we have specified the following dimension: *importance for the adaptive web*. This is intended to help the researcher in the choice of the right technique for a specific situation by summarizing the way in which the method could be especially helpful for adaptive web-based systems.

At the end of every section we have also added a table providing criteria which should be helpful in choosing the most appropriate method to be applied in respect to that particular temporal phase presented in the corresponding section. For these purposes the table classifies the methods according to the following dimensions:

- *Kind of factors*, which highlights the factors the methods are most suited to generate and evaluate.
- *Applicability conditions*, which underline if there are constraints or particular conditions necessary to utilize methodologies.
- *Pros and cons*, which summarize advantages and disadvantages deriving from the application of each method.

1.2.2 Data Collection Methods

Before presenting methods and techniques for evaluation, it is worth describing data collection methods and how they interact together. Evaluation experts can choose between different methods and data collection tools depending on a number of circumstances (e.g., the type of evaluation techniques, the temporal phase, eventual constraints, etc). It is possible, in connection with a particular evaluation technique, to use more than one data collection method (e.g., users can be observed in a controlled experiment and queried at the end by means of a questionnaire). The data collection methods will be examined below.

The collection of user's opinion.

The collection of user's opinion, also known as *query technique*, is a method that can be used to elicit details about the user's point of view of a system and it can be particularly useful for adaptive web systems in order to collect ideas and details to produce adaptation.

Questionnaires.

Questionnaires have pre-defined questions and a set of closed or open answers. The styles of questions can be general, open-ended, scalar, multi-choice, ranked. Questionnaires are less flexible than interviews, but can be administered more easily (for details see [43]). Questionnaires can be used to collect information useful to define the knowledge base of the system for user modeling or system adaptations (especially in the requirement phase, see Section 1.3.1). For instance, questionnaires and scenarios³ have been used for creating a user modeling component [1].

Since a large number of responses to a questionnaire is required in order to generalize results (which, otherwise, could be biased), existing surveys about the target population (e.g., psycho-graphic and lifestyle surveys, web-users researches, etc) can be exploited for the knowledge base definition, to build stereotype-based user-modeling systems (see for example [49] and [61]), or to inspire the adaptation strategies (see for instance Chapter 16 of this book [57]). Questionnaires (and log files) can also be used to evaluate the accuracy of system recommendations.

In adaptive web systems and their evaluation, questionnaires can further be exploited as:

- *on-line questionnaires*, to collect general users' data and preferences in order to generate recommendations. For instance, they can be used to acquire

³ A scenario is aimed at illustrating usage situations by showing step-by-step the possible user's actions and options. It can be represented by textual descriptions, images, videos and it can be employed in different design phases

a user interest profile in collaborative [136] and feature-based recommender systems (see [126] and Chapter 18 of this book [118]). At the beginning, the system can use the user's rating to generate recommendations. Then, the data collected through the questionnaires (and web log files) can also be used to evaluate the accuracy of system recommendations by comparing the system assumptions with the real user choices [35, 102, 5].

- *pre-test questionnaires*, to establish the user's background and place her within the population of interest, and/or to use this information to find a possible correlation after the test session (e.g., computer skilled users could perform better, etc). Pre-test questionnaires can also be useful to gather data in order to classify the user before the experimental session (for instance in a stereotype [5]).
- *post-test questionnaires*, to collect structured information after the experimental session, or after having tried a system for a while. For instance, Matsuo [98] asked the users questions about the system functionality using a 5-point Likert scale⁴, Alfonseca and Rodrguezi [2] asked questions concerning usability of their system, Bul [26] used a post-test questionnaire about the potential utility of their system. Besides, post-test questionnaires can be exploited to compare the assumption in the user model to an external test [158].
- *pre and post-test questionnaires*, exploited together to collect changes due to real or experimental user-system interaction. For instance, in adaptive web-learning systems, pre and post-test questionnaires can be exploited to register improvements in the student's knowledge after one or more interactions. Pre-test questionnaires can also be used to group students on the basis of their ability [103], their knowledge [138], their motivational factors and their learning strategies [70] and then to test separately the results of the different groups (with post-test questionnaires), or to propose to the different groups solutions adapted to their cognitive profile [60].

Interviews.

Interviews are used to collect self-reported opinions and experiences, preferences and behavioral motivations [43, 109]. Interviews are more flexible than questionnaires and they are well suited for exploratory studies (see for instance contextual design, Section 1.3.1). Interviews can be structured, semi-structured, and unstructured.

Structured interviews have been exploited in combination with scenarios to identify adaptivity requirements [157]. However, in this experiment, results were not satisfactory to their purpose and they suggested alternative approaches to elicit requirements, such as mock-up prototypes. Unstructured interviews are often used after a test session to gather user's opinion, such as the user's satisfaction with the system [53].

⁴ A Likert scale is a type of survey question where users are asked to evaluate the level at which they agree or disagree with a given sentence

User observation methods.

This family of methods is based on direct or indirect user's observation. They can be carried out with or without predetermined tasks.

Think aloud protocols.

Think aloud protocols are methods that make use of the user's thought throughout the experimental session, or simply while the user is performing a task. In think aloud protocols the user is explicitly asked to think out loud when she is performing a task in order to record her spontaneous reactions. The main disadvantage of this method is that it disturbs performance measurements. See for example [121] who have encouraged their users to think aloud while performing experimental tasks for the evaluation of a user modeling system based on the theory of information scent. Another possible protocol is **constructive interaction**, where more users work collaboratively to solve problems at the interface.

User observation.

Observation is a data collection method wherein the user's behavior is observed during an experimental session or in her real environment when she interacts with the system. In the former case, the user's actions are usually quantitatively analyzed and measurements are taken, while in the latter case the user's performance is typically studied from a qualitative ⁵ point of view. Moreover, as described in Chapter 17 of this book [84] about the evaluation of the GUIDE system, a user study can be based at the same time on direct observation, audio recording and logging data.

⁵ "The choice between quantitative and qualitative methodologies depends on the point of view of the evaluation: while quantitative research tries to explain the variance of the dependent variable(s) generated through the manipulation of independent variable(s) (variable-based), in qualitative research the object of the study becomes the individual subject (case-based). Qualitative researchers sustain that a subject cannot be reduced to a sum of variables and therefore a deeper knowledge of a fewer group of subjects is more useful than an empirical experiment with a representative sample. Even if the final goals of both approaches are similar (they both want to come up with predictive theories to generalize over individual behaviours), they are carried out in a different way: while quantitative researchers try to explain the cause-effect relationships between variables and make generalizations on the obtained results (extensive approach), qualitative researchers want to comprehend the subjects under the study by interpreting their points of view and by analyzing the facts in depth (intensive approach) in order to propose new general understanding of the reality." [51]

Logging use.

The logging use can be considered a kind of indirect observation and consists in the analysis of log files that register all the actions of the users. The log files analysis shows the real behavior of users and is one of the most reliable ways to demonstrate the real effectiveness of user modeling and adaptive solutions [9, 25, 41, 87, 137].

Log data have been used to run algorithm with real users data [10, 133], and for simulations such as reproducing Web surfing [85], simulating e-mail usage [101], and calculating the accuracy of the system's predictions. Log file analysis can also suggest the way to design system adaptation on the basis of the behavior of the users (see also "Automatic usability testing and web usage mining" in Sec. 1.3.3). For instance, Herder et al. [63] conducted a long-term client study to investigate the design implication for more personalized browser history support.

Table 1. Data Collection Methods

	Kind of Factors	Applicability conditions	Pros and cons
Data Collection Methods			
Questionnaires and surveys	Demographic data, users' opinions, preferences and attitudes	A sample of target users	+ users involvement; subjective data - data may be biased by non representative sample; questions must be phrased carefully
On-line Questionnaires	User opinions, user satisfaction, user preferences	A sample of target users; web application	+ users involvement; subjective data - data may be biased by non representative sample; questions must be phrased carefully; little control over participation
Pre-test Questionnaires	Classification of users	A sample of target users; an adaptive web-based prototype/system to be tested	+ external user classification - erroneous classification
Post-test Questionnaires	User opinions, user satisfaction	A sample of target users; an adaptive web-based prototype/system to be tested	+ subjective data - data may be biased
Pre and post-test Questionnaires	Learning gain, change in opinion or attitude	A sample of target users; an adaptive educational web-based prototype/system to be tested	+ measuring change or development - sequential effects

	Kind of Factors	Applicability conditions	Pros and cons
Interviews	User opinion, user satisfaction	A sample of target users	+ subjective data - time consuming; subjective interpretation
Think aloud protocol	Cognitions; usability of interface adaptations	A sample of target users; an adaptive web-based prototype/system	+ user's spontaneous reactions; information about what the users are thinking/why they do something - interferes with human thought processes, so actions cannot be measured; protocol might bias actions and performance
Users observation	Observation of real user-system interactions	A sample of target users; an (adaptive) web-based prototype/system	+ observing user in action - user may be influenced
Logging use	Real usage data, data for simulation, clustering	A sample of target users; an adaptive web-based prototype/system to be tested	+ large amount of useful information; unobtrusive - many client-side actions not recorded; no information about what the users are thinking/why they do something

1.3 Phases of Evaluation

The techniques described in this section can be categorized according to the phases of the development cycle they are usually used in, i.e., the requirement phase, the preliminary evaluation phase, and the final evaluation phase.

1.3.1 The requirement phase

The requirement phase is usually the first phase in the system design process. It can be defined as a “process of finding out what a client (or a customer) requires from a software system” [125]. During this phase it can be useful to gather data about typical users (features, behavior, actions, needs, environment, etc), the application domain, the system features and goals, etc.

In the case of adaptive web-based systems, the choice of relevant features to model the user (such as goals and plans of the user, the social and physical environment, etc) and consequently adapt the system, may be aided by prior knowledge of the real users of the system, the context of use, and domain experts’ opinion. A deeper knowledge of these factors can offer a broader view of the application goals and prevent serious mistakes, especially in the case of innovative systems. As Benyon [17] has underlined, adaptive systems should benefit more than other systems from the requirement analysis before starting any kind of evaluation, because a higher number of features has to be taken into account in the development of these systems. The recognition that an adaptive capability may be desirable leads to a improved system analysis and design.

According to Benyon, five related and interdependent activities need to be considered in the requirement phase of an adaptive system:

- *functional analysis*, aimed at establishing the main functions of the system;
- *data analysis*, concerned with understanding and representing the meaning and structure of data in the application;
- *task knowledge analysis*, focused on the cognitive characteristics required by users of the system such as the user’s mental model, cognitive loading, the search strategy required, etc.;
- *user analysis*, that determines the scope of the user population that the system is to respond to. This is concerned with obtaining attributes of users that are relevant for the application such as required intellectual capability, cognitive processing ability, and similar. The target population will be analyzed and classified according to the aspects of the application derived from the point mentioned above;
- *environment analysis*, that covers the environment within which the system is to operate.

The above activities presented by Benyon directly correspond to the following stages of the requirement analysis [125]. In the following we present techniques for gathering requirements highlighting the specific contribution for adaptive web systems, according to Benyon’s proposal.

Task analysis.

Task analysis methods are based on breaking down the tasks of potential users into users' actions and users' cognitive processes [43]. In most cases, the tasks to be analyzed are broken down into sub-tasks (see for instance *Hierarchical Task Analysis* (HTA), [40]). So far, there has been little experience in the application of this method to adaptive web-based systems, even if task analysis could be used to deeply investigate users' actions and plans in order to decide in advance in which phase of the interaction the system could propose adaptations. For instance, if the task analysis shows that a goal can be reached faster by proposing some shortcut in the interface, adaptation can be proposed at that point in order to anticipate the user's plans. Task analysis results can also be useful to avoid the well-known cold-start problem⁶ of knowledge-based systems by proposing default adaptations at the beginning of the user-system interaction. For instance, if it is possible to identify different kinds of target users of the website (e.g., students, teachers, administration staff, etc), task analysis can investigate the main goals of these typical users (e.g., students want to check course timetables and examination results, teachers want to insert course slides, etc), analyze in depth the tasks to be performed, and propose possible adaptations.

Importance for the adaptive web: useful for functional, data, and task knowledge analysis of Benyon's classification.

Cognitive and socio-technical models.

The understanding of the internal cognitive process as a person performs a task, and the representation of knowledge that she needs to do that, is the purpose of the cognitive task models [43, 125]. An example of goal-oriented cognitive model is the GOMS model (Goals, Operators, Methods and Selection) that consists of descriptions of the methods (series of steps consisting of actions performed by the users) needed to accomplish specific goals. For instance, cognitive models have been applied in the development of a mixed-initiative framework [27], by investigating the performance implications of customization decisions by means of a simplified form of GOMS analysis.

Additional methods for requirements analysis also include socio-technical models, which consider social and technical issues and recognize that technology is a part of a wider organizational environment [43]. For instance, the USTM/ CUSTOM [81] model focuses on establishing stakeholder requirements⁷. Even if seldom applied in the adaptive web, both goal-oriented cogni-

⁶ Adaptive web-based systems can suffer from cold-start problem, when no initial information about the user is available early on upon which to base adaptations.

⁷ A stakeholder is here defined as anyone who is affected by the success or the failure of the system (e.g., who uses the systems, who receive output from it or provide input, etc) [43].

tive models and socio-technical models could offer fruitful contributions during the design phase since they are strong generative models [43]. They can help to make predictions respectively about the internal cognitive processes and the social behaviors of users, and therefore be adopted in the construction of the user model knowledge base and the corresponding system adaptations.

Importance for the adaptive web: useful for task knowledge and user analysis of Benyon’s classification.

Contextual design.

Contextual design is usually organized as a semi-structured interview (see Sec. 1.2.2) covering the interesting aspects of a system while users are working in their natural work environment on their own work [18, 125]. Often the interview is recorded in order to be elaborated on by both the interviewer and by the interviewee ⁸. Contextual design is a qualitative observational methodology that can be applied in the adaptive web in order to gather social and environmental information (such as structure and language used at work; individual and group actions and intentions; the culture affecting the work; explicit and implicit aspects of the work, etc) useful to inspire the design of system adaptations.

Contextual design has been used in Intelligent Tutoring Systems, for instance, through the observations of the strategies employed by teachers [3]. Masthoff [97] has also exploited contextual design together with a variant of Wizard of Oz studies.

Importance for the adaptive web: useful for user and environment analysis of Benyon’s classification.

Focus group.

Focus group [58], [109] is an informal technique that can be used to collect user opinions. It is structured as a discussion about specific topics moderated by a trained group leader [58]. A typical focus group session includes from 8 to 12 target users and lasts around for two hours.

Depending on the type of users involved (e.g., final users, domain experts, technicians) focus groups can be exploited to gather functional requirements, data requirements, usability requirements, and environmental requirements to be considered in the design of system adaptations. For instance, during the development of an adaptive web-based system for the local public administration, mock-ups have been developed which had been discussed and redesigned

⁸ The additional “testing” after-the-fact is also known as **retrospective testing**, and it is usually conducted after a recorded user testing session. Retrospective testing consist in reviewing the tape with the user to ask additional questions and get further clarification.

after several focus group sessions with experts and final users involved in the project [52]. Focus groups can also be successfully used in combination with other methods that simulate the interaction phases when the system is not yet implemented. For instance, van Barneveld and van Setten [149] use focus groups and creative brainstorming sessions to inspire a recommender systems user interface.

Importance for the adaptive web: useful for functional, data, user and environment analysis of Benyon's classification.

Systematic observation.

Systematic observation can be defined as a "particular approach to quantifying behavior. This approach is typically concerned with naturally occurring behavior observed in a real context" [6]. The observation is conducted in two phases: First, various forms of behavior, so-called behavioral codes are defined. Secondly, observers are asked to record whenever behavior corresponding to predefined codes occurs. The data can be analysed in two ways: either in the form of non-sequential analysis (subjects are observed for the given time slots during different time intervals) or as sequential analysis (subjects are observed for a given period of time).

In the adaptive web, systematic observation can be used during the requirement phase to systematically analyze significant interactions in order to discover interaction patterns, recurrent and typical behavior, the user's plans (e.g., sequences of user actions-interactions, distribution of user's activities along the time, etc) that can be modelled by the adaptation. For instance, in order to model teaching strategies in an Intelligent Tutoring System, Rizzo et al. [128] recorded the interactions taking place between the tutor and the student in a natural setting or computer-mediated interface. Then the records were systematically analyzed to find teaching patterns useful to inspire adaptation mechanisms.

Importance for the adaptive web: useful for task knowledge, user and environment analysis of Benyon's classification.

Table 2. Requirements phase

	Kind of Factors	Applicability conditions	Pros and cons
Requirement phase			
Task analysis	Fine decomposition of user actions; focus on cognitive processes required by the users	Formalization of the system; presence of expert evaluators	+ find out where adaptation is required – possibly artificial situations
Cognitive models	Discovering of cognitive factors to be considered during the design of system adaptations	Formalization of the system; presence of expert evaluators	+ consideration of cognitive factors in the early phases – expert evaluators required
Socio-technical models	Discovering social factors to be considered during the design of system adaptations	Representative users collaboration; presence of expert evaluators	+ consideration of social factors in early phases – expert evaluators required
Contextual design	Social and environmental information (such as structure and language used at work; individual and group actions and intentions; the culture affecting the work; explicit and implicit aspects of the work, etc)	Existing or similar web sites; representative users collaboration	+ valuable insights into usage context – expert in qualitative evaluation required

	Kind of Factors	Applicability conditions	Pros and cons
Focus group	Gathering of heterogeneous requirement data from real users and domain experts for the design of system adaptations	Contact with target and representative users, technicians, domain experts	<ul style="list-style-type: none"> + final users involvement; gathering of users opinions and requirements during an informal discussion - not to be confused with final evaluations
Systematic observation	Systematic analysis of features (e.g., interaction patterns, recurring activities, etc) that can be modeled by adaptation strategies	Contact with target and representative users; existing or similar web sites	<ul style="list-style-type: none"> + quantification of user activities; observation of users in their natural context - expensive; expert in qualitative evaluation/observation required; time consuming

1.3.2 Preliminary evaluation phase

The preliminary evaluation phase occurs during the system development. It is very important to carry out one or more evaluations during this phase to avoid expensive and complex re-design of the system once it is finished. It can be based on predictive or formative methods.

Predictive evaluations are aimed at making predictions, based on experts' judgement, about the performance of the interactive systems and preventing errors without performing empirical evaluations with users. Formative evaluations are aimed at checking the first design choices before actual implementation and getting the clues for revising the design in an iterative design-re-design process.

Heuristic evaluation.

A heuristic is a general principle or a rule of thumb that can guide a design decision or be used to critique existing decisions. Heuristic evaluation [113] describes a method in which a small set of evaluators examine a user interface and look for problems that violate some of the general principles of good interface design.

Unfortunately, in the adaptive web field a set of recognized and accepted guidelines to follow is still missing. On the one side, this lack can be filled only by publishing statistically significant results that can demonstrate, for instance, that one adaptation strategy is better than another one in a given situation, or that some adaptation technique should be carefully applied. For instance, Sears & Shneiderman [134] performed an evaluation on menu choices sorted on the basis of their usage frequency. Their results reported that the users were disoriented by the menu choices sorted on usage frequency because of the lack of order in the adapted menu. A preferable solution could be the positioning of the most often used choices at the top of the list before all the other ordered items (the so-called *split menu*). Therefore, researchers should be careful in applying this technique. The key point is to carry out evaluations leading to significant results that can be re-used in other research, and promote the development of standard measures that would be able to reasonably evaluate the systems' reliability. To this purpose, Weibelzahl & Weber [156] promoted the development of an online database for studies of empirical evaluations to assist researchers in the evaluation of adaptive systems and to promote the construction of a corpus of guidelines.

On the other side, also general principles have to be considered. For instance, Magoulas et al. [93] proposed an integration of heuristic evaluation in the evaluation of adaptive learning environments. They modified the Nielsen's heuristics [109] to reflect pedagogical consideration and then they collocated their heuristics into the level of adaptation [93]. E.g., the Nielsen's heuristic "Recognition rather than recall" is specified in "instructions and cues that the system provides for users to identify results of adaptations easily". As

sketched in Section 1.3.3, Jameson [73] proposed five usability challenges for adaptive interfaces to deal with usability problems that can arise with these systems.

Importance for the adaptive web: making prediction about the usability and the applicability of interface adaptations.

Domain Expert review.

In the first implementation phases of an adaptive web site, the presence of domain experts and human designers can be beneficial. For instance, a domain expert can help defining the dimensions of the user model and domain-relevant features. They can also contribute towards the evaluation of correctness of the inference mechanism [5] and interface adaptations [54]. For instance, an adaptive web site that suggests TV programs can benefit from audience TV experts working in TV advertising that may illustrate habits, behaviors and preferences of homogeneous groups of TV viewers. In this specific case a domain expert review can be beneficial in the requirement phase.

For example, Chapter 1 outlines how experts can contribute to the development of an uncertainty-based user model [23]. Experts can also be asked to pick up a set of relevant documents for a certain query and their judgments are used to check the correctness of system recommendations. For examples of evaluation of a recommender system with the estimation of precision and recall returned to a human advisor proposal see [92]. More metrics for evaluating recommender systems without users are listed in Chapter 3 [105].

Expert review, as well as cognitive walkthrough, scenario-based design and prototypes, can be used to evaluate *parallel designs* [109], which consist of exploring different design alternatives before setting on a single proposal to be developed further. Parallel design can very suitable for systems that have a user model since in this way designers can propose different solutions (what to model) and different interaction strategies (what the user can control) depending on the identified users. Parallel design is a very useful approach since it lets one to explore adaptive solutions and simulate strategies with users before the system is implemented. Design rationale⁹ and design space analysis¹⁰ can also be helpful in context of exploring and reasoning among different design alternatives. For details about design rationale see [90], while for design space analysis see [15]. Experts can be involved in **coaching methods**, which are usability testing techniques wherein users are encouraged to ask questions to the expert/coach, who responds with appropriate instruction. Typical user

⁹ Design rationale “is the information that explains why a computer systems is the way it is, including its structural or architectural description and its functional or behavioral description” [43].

¹⁰ Design space analysis is an “approach to design that encourages the designer to explore alternative design solution” [125]

questions help at identifying usability problems.

Importance for the adaptive web: predicting the correctness of inference mechanisms and usability of interface adaptations; simulations of design alternatives.

Cognitive walkthrough.

Cognitive walkthrough is an evaluation method wherein experts play the role of users in order to identify usability problems [124]. Similar to heuristic evaluation, this predictive technique should benefit from a set of guidelines for the adaptive web that should help evaluators to assess not only general HCI mistakes but also recognized errors in the design of adaptations.

Importance for the adaptive web: making prediction about the usability and the reliability of interface adaptations that help the user to accomplish tasks.

Wizard of Oz prototyping.

Wizard of Oz prototyping [109, 125] is a form of prototyping in which the user appears to be interacting with the software when, in fact, the input is transmitted to the wizard (the experimenter) who is responding to user's actions. The user interacts with the emulated system without being aware of the trick.

Wizard of Oz prototyping can be applied in the evaluation of adaptive web systems, for instance, when a real time user-system interaction has to be simulated in the early implementation phases (e.g., speech recognition, interaction with animated agents, etc). For example, a Wizard of Oz interface that enables the tutor to communicate with the student in a computer-mediated environment has been used to model tutorial strategies [128]. Maulsby, Greenberg & Mander [99] used Wizard of Oz to prototype an intelligent agent, and Masthoff [97] applied a variant of Wizard of Oz under a contextual design point of view, making users to take the role of the wizard: humans tend to be good at adaptation, thus, observing them in the role of the wizard may help to design the adaptation.

Importance for the adaptive web: simulation of a real time user-adapted interaction.

Prototyping.

Prototypes are artifacts that simulate or animate some but not all features of the intended system [43]. They can be divided in two main categories: static, paper-based prototypes and interactive, software-based prototypes. Testing

prototypes is very common, however they should not be considered to be finished products. Prototypes can also be: *horizontal*, when they contain a shallow layer of the whole surface of the user interface; *vertical*, when they include a small number of deep paths through the interface, but do not include any part of the remaining paths; *scenario-based* when they fully implement some important tasks that cut through the functionality of the prototype. For instance, Gena & Ardissono [52] evaluated an adaptive web prototype in a controlled experiment with real users. The main aims of the test were to discover whether the interface adaptations were visible and effective and whether the content adaptations were consistent and helpful to the task completion. In Chapter 17 of this book [84] is reported a prototype evaluation of the TellMaris system.

As described above for parallel design, scenario based prototypes can be helpful at simulating adaptation strategies and design alternatives with real users and expert before the initial implementations.

Importance for the adaptive web: early evaluation of adaptation strategies; simulations of adaptations strategies and design alternatives.

Card sorting.

Card sorting is a generative method for exploring how people group items and it is particularly useful for defining web site structures [129]. It can be used to discover the latent structure of an unsorted list of categories or ideas. The investigator writes each category on a small index card (e.g., the menu items of a web site), and requests users to groups these cards into clusters (e.g., the main item of the navigational structure). The clusters can be predefined (closed card sorting) or defined by the user herself (open card sorting). So far, there has been little experience of card sorting in adaptive web systems. Card sorting could be carried out with different groups of representative users for the definition of the information architecture of an adaptive web site. It can inspire different information structures for different groups of users (e.g., how novice and experts see the structure of the web site information space).

Importance for the adaptive web: definition of different information architectures for different group of representative users.

Cooperative evaluation.

An additional methodology that can be carried out during the preliminary evaluation phase is the cooperative evaluation [107], which includes methods wherein the user is encouraged to act as a collaborator in the evaluation to identify usability problems and their solutions. Even if seldom applied, cooperative evaluation is a qualitative technique that could be applied in the evaluation of adaptive web based systems to detect general problems (e.g.,

usability, reliability of adaptations, etc) in early development phases and to explore the user's point of view to collect design inspiration for the adaptive solutions.

Importance for the adaptive web: detection of general problems concerning adaptations; design inspirations for adaptive solutions.

Participative evaluation.

Another qualitative technique useful in the former evaluation phases is the participative evaluation [109, 125] wherein final users are involved with the design team and participate in design decisions. Participative evaluation is strictly tied to participatory design techniques where users are involved in all the design phases [58, 59]. So far, this methodology is rather disregarded in the adaptive web, however it could be applied to have users directly participating at the design of adaptation strategies.

Importance for the adaptive web: gathering of heterogenous requirement data from real users and domain experts; users and expert participating at the design of adaptation strategies.

Table 3. Preliminary evaluation phase

	Kind of Factors	Applicability conditions	Pros and cons
Preliminary evaluation phase			
Heuristic evaluation	Usability of interface adaptations	A user-adaptive prototype/system	+ making prediction about the interface design without involving users - guidelines for adaptive systems are still missing
Expert review	User, domain and interface knowledge	Only for (adaptive web) knowledge-based system	+ valuable source of information for the system KB - experts may use background and contextual knowledge that are not available to a system
Cognitive walkthrough	Usability of interface adaptations	A user-adaptive prototype/system; presence of expert evaluators	+ make prediction about the design without involving users - guidelines for adaptive systems are still missing; time intensive
Wizard of Oz simulation	Early prototype evaluation	Only for systems that simulate a real time user-system interaction	+ useful when the system is still not completed - the Oz is more intelligent than the system!
Prototypes	Evaluation of vertical or horizontal prototype; design of scenarios	A running user-adaptive prototype	+ evaluation in early phases; simulation of scenarios - limited functionality available

	Kind of Factors	Applicability conditions	Pros and cons
Card sorting	To set the web site structure from the user's point of view	<ul style="list-style-type: none"> - Top-down information architecture of the web site to be adapted 	<ul style="list-style-type: none"> + considers the user's mental model - only applicable to web sites that have a categories-based information architecture (top-down)
Cooperative evaluation	Detection of general problems concerning adaptations in the early stages	Final users collaborating during the design-redesign phase	<ul style="list-style-type: none"> + direct, immediate user feedback and suggestions - not all the target users are considered
Participative evaluation	Gathering of heterogeneous requirement data from real users and domain experts	Final users involved in the design team	<ul style="list-style-type: none"> + direct, immediate user suggestions and inspirations - not all the target users are considered

1.3.3 Final evaluation phase

The final evaluation phase occurs at the end of the system development and it is aimed at evaluating the overall quality of a system with users performing real tasks.

Usability testing.

According to the ISO definition ISO 9241-11:1998 usability is “the extent to which a product can be used by specified users, to achieve specified goals, with effectiveness, efficiency and satisfaction, in a specified context of use” [68]. Based on this definition, the usability of a web site could be measured by how easily and effectively a specific user can browse the web site, to carry out a fixed set of tasks, in a defined set of environments [31].

The core of usability testing [109], [130], [44] is to make the users use the web site and record what happens. In this way it is possible to evaluate the response of a real user rather than to propose interfaces as designed by the designers. In particular, the usability test has four necessary features:

- participants represent real users;
- participants do real tasks;
- users’ performances are observed and sometimes recorded (see Sec. 1.2.2);
- users’ opinions are collected by means of interviews or questionnaires (see Sec. 1.2.2).

According to [110] a usability test with 4-5 representative users will discover 80% of major usability problems of a web site, while 10 users will discover up to 90% of problems.

One or more usability tests on an adaptive web site should always be performed. The usability of adaptive interfaces has been widely discussed, this will be reported in Section 1.4. Due to inherent problems tied to adaptive interfaces and to the importance of usability in the web, the usability of an adaptive web site should always be tested by taking into account both interface adaptations and general interface solutions. Some examples of usability testing in the adaptive web can be found in [2, 16, 131, 133], while for details on testing procedures see [44, 109, 130].

Jameson [74] pointed out that the anticipation and the prevention of usability side effects should form an essential part of the iterative design of user-adaptive systems. Jameson [73] proposed five usability challenges for adaptive interfaces: (1) predictability and transparency, (2) controllability, (3) unobtrusiveness, (4) privacy, and (5) breadth of experience. He tried to match usability goals and typical adaptive systems properties to deal with usability problems which these systems can suffer. Transparency and controllability, nevertheless, could imply involving the user in the personalization process and/or adding some adaptability into the system. But sometimes users have difficulty understanding and controlling personalization. For an evaluation of

problems with transparency and control of adaptive web systems see [72], [38]. However, there are also examples of learning systems that show systems that expose the user model to the student enhance learning [78],[104]. It is important to notice that usability tests of adaptive web sites can only be applied to evaluate general usability problems at the interface. If one would test the usability of one adaptation technique compared to another one, a controlled experiment should be carried out.

Importance for the adaptive web: usability of the overall web site and of interface adaptations.

Automatic usability testing and web usage mining.

In recent years, interest in automatic tools able to support the evaluation process has been increasing. The methods for usability evaluation of Web sites has been classified into two types of approaches [117]: *methods based on empirical evaluation*, where user's logs data generated by a web server are analyzed, and *methods based on analytical evaluation*, where various combinations of criteria, guidelines and models are automatically applied.

In the former ones, the analysis of real usage data is considered to be a solution to discover real user-system interaction. For instance, Web usage analysis [106, 139, 120] is a long process of learning to see a website from the perspective of its users. By analyzing Web server log data usage patterns could be discovered (e.g., pages occurring frequently together and in the same order). This may be a signal that many users navigate differently than originally anticipated when the site was designed. The usage mining process can involve the discovery of association rules, sequential patterns, page view clusters, user clusters, or any other pattern discovery methods. After having collected web log data and reached some evidence (confirmed by statistical analysis), the re-design of the interface may be accomplished in two ways [119]:

- by *transformation*, improving the site structure based on interactions with all visitors.
- by *customization*, adapting the site presentation to the needs of each individual visitor based on information about those individuals.

Between these two alternatives, a third solution could be adopted: personalizing a site according to a different cluster of users' behavior (for instance occasional, regular, novice, expert user, etc) emerged from the data mining process. Finally, to help the analysis of this large amount of data, logs of user interactions can be analyzed through graphical tools that visualize the paths followed by the users during the site visit [37].

Analytical methods comprehend automatic tools such as Bobby¹¹, that verifies the application of accessibility guidelines; WebSat¹², that evaluates

¹¹ <http://www.cast.org/bobby>

¹² <http://www.research.att.com/conf/hfweb/proceedings/scholtz/index.html>

usability by analyzing the HTML code through the application of usability guidelines; or Design Advisor¹³, which is based on eye-tracking techniques.

Between analytical and empirical methods are mixed approaches that combine the analysis of browsers logs with usability guidelines and models of user's actions. See for example [117].

Importance for the adaptive web: usability of the overall web site and of interface adaptations; inspiration for the adaptive behavior of the web site.

Accessibility.

According to the ISO definition ISO/TS 16071:2003 accessibility is “the usability of a product, service, environment or facility by people with the widest range of capabilities” [69]. This definition strictly correlates accessibility to usability, with the difference that an accessible web site must be usable for every one, also for people with disabilities. There are a variety of tools and approaches for evaluating Web site accessibility. For more details see [150].

Adaptive web sites, which by definition pay more attention to users' needs, should respect accessibility guidelines. Moreover, they could adapt to the specific users with disabilities taking into account their specific problems, since impaired users need their specific requirement. For example, in the AVANTI project, the system adapted the content and the presentation of web pages to each individual user, also taking into account elderly and disabled users [47]. Stephanidis [141] highlighted the potential adaptive techniques have to facilitate both accessibility and high quality interaction, for the broadest possible end-user population.

Importance for the adaptive web: proposing adaptive solutions for different groups of disabled users to increase the accessibility of the web site.

Controlled experiments.

Controlled experiments [79, 80] are one of the most relevant evaluation techniques for the development of the adaptive web, and their impact in user-adapted systems has been largely discussed [32, 51]. Indeed, they are often performed in the evaluation of adaptive systems (mostly for the evaluation of interface adaptations), but sometimes experiments are not properly designed and thus they do not produce significant results to be taken into account. As will be discussed in Section 1.4 significant results are necessary for the growth of the adaptive web, because they can be extended to provide generalizations and guidelines for future works, therefore it is important to correctly carry out every design step and evaluate results with the required statistics.

¹³ <http://www.tri.sbc.com/hfweb/faraday/faraday.htm>

The general idea underlying a controlled experiment is that by changing one element (the independent variable) in a controlled environment its effects on user's behavior can be measured (on the dependent variable). The aim of a controlled experiment is to empirically support a hypothesis and to verify cause-effect relationships by controlling the experimental variables. Therefore, as described in [73], controlled experiments can be used to evaluate the accuracy of modeling (content layer: e.g. are the system recommendations correct?) and the usability of the adaptive system (interface layer: e.g. do the interface adaptations enhance the quality of the interaction?). The most important criteria to follow in every experiment are:

- participants have to be credible: they have to be real users of the application under evaluation;
- experimental tasks have to be credible: users have to perform tasks usually performed when they are using the application;
- participants have to be observed during the experiment (see Sec. 1.2.2) and their performance recorded;
- finally, users' opinions are collected by means of interviews or questionnaires (see Sec. 1.2.2).

Empirical evaluation takes place in a laboratory environment. Well equipped laboratory may contain sophisticated audio/video recording facilities, two-way mirrors, and instrumented computers. On the one hand, the lack of context, and the unnatural condition creates an artificial situation, far from the place where the real action takes place. On the other hand, there are some situations where the laboratory observation is the only option, for instance if the location is dangerous and sometimes the experimenters may want to deliberately manipulate the context in order to create unexplored situations [Dix et al. 1998]. The schematic process of a controlled experiment can be summarized in the following steps [80], while a more detailed discussion on problems that can arise will be presented in Sec. 1.4.

Develop research hypothesis.

In statistics, usually two hypotheses are considered: the null hypothesis and the alternative hypothesis. The null hypothesis foresees no dependencies between independent and dependent variables and therefore no relationships in the population of interest (e.g., the adaptivity does not cause any effect on user performance). On the contrary, the alternative hypothesis states a dependency between independent and dependent variables: the manipulation of the independent variable(s) causes effects on the dependent variable(s) (e.g., the adaptivity causes some effects on user performance).

Identify the experimental variables.

The hypothesis can be verified by manipulating and measuring variables in a controlled situation. In a controlled experiment two kinds of variables can

be identified: independent variable(s) (e.g., the presence of adaptive behavior in a web site) and dependent variable(s) (e.g., the task completion time, the number of errors, proportion/qualities of tasks achieved, interaction patterns, learning time/rate, user satisfaction, number of clicks, back button usage, home page visit, cognitive load measured through blood pressure, pupil dilatation, eye-tracking, number of fixations and fixation times, etc). See [75] for an example of how these variables are measured and analyzed during an evaluation of an adaptive web-based system; [13], [34] for eye-tracking in user modeling systems, and [67] for an experimental methodology to evaluate cognitive load in adaptive information filtering.

It is important to notice that it could also be interesting to analyze the correlation between variables that are characteristics naturally occurring in the subject. Statistical correlation (for more details see [79]) tells whether there is a relationship between two variables. In this kind of experiments, namely *correlational studies*, both variables are measured because there are no true independent variables. For example [66] an empirical study of adaptive help system for web-based applications correlated the ACT-value of procedural knowledge with subjective and objective measures of performance. For other examples of correlational studies see [95].

Select the subjects.

The goal of sampling is to collect data from a representative sample drawn from a larger population to make inferences about that population. A common problem of most evaluations in adaptive systems is that often the sample is too narrow to produce significant results. Rules of thumb for the sampling strategies are: i) the number of subjects has to be representative of the target population, ii) they should fit the statistics applied in data analysis, iii) they should fit subjects and resources availability.

Select the experimental methods and conduct the experiment.

The selection of an experimental method consists primarily of collecting the data using a particular experimental design. The simplest design for an experiment is the **single factor design** in which one independent variable is manipulated (e.g., is the adaptive version more successful or the one without adaptations?). When two or more independent variables are manipulated the design is called **factorial design** (e.g., testing the adaptivity and the scrutability of an adaptive web site). Then, subjects are assigned to different treatment conditions. In the simplest procedure, the **between-subjects design**, an experimental group of subjects is assigned to the treatment (e.g., adaptivity), while another group of subjects, the control group, is assigned to the condition consisting of absence of a specific experimental treatment. For example in [91], six users conducted dialogs with the adaptive version of system, and six other users conducted dialogs with the non-adaptive one; while Farzan & Brusilovsky [46] have evaluated a course recommendation system

by preparing two different version of the system: one with social navigation support (experimental group) and the other one without (control group).

There may be more than two groups, depending on the number of independent variables and the number of levels each variable can assume.

At the other extreme is the **within-subjects design** in which each subject is assigned to all treatment conditions (e.g., subjects completing tasks using both the application with adaptations and the one without). For instance, in the evaluation of a learning system that adapts the interface to the user's cognitive style, the same subjects used the system under three different treatment conditions [147]. Kumar [86] proposed a within-subject approach categorizing student-concepts as control and test groups instead of the student themselves. In between are designs in which the subjects are serving in some but not all the treatment conditions (**partial, or mixed, within-subjects factorial design**). For example, in [50] the subjects were split into two groups and every group completed the tasks with and without system adaptations (the tasks completed without adaptations by one group were completed with adaptations by the other one, and vice versa).

In an ideal experiment only the independent variable should vary from condition to condition. In reality, other factors are found to vary along with the treatment differences. These unwanted factors are called **confounding variables** (or nuisance variables) and they usually pose serious problems if they influence the behavior under study since it becomes hard to distinguish between the effects of the manipulated variable and the effects due to confounding variables. As indicated by [32], one way to control the potential source of confounding variables is holding them constant, so that they have the same influence on each of the treatment conditions (for instance, the testing environment, the location of the experiment, the instructions given to the participants may be controlled by holding them physically constant). Unfortunately, not all the potential variables can be handled in this way (for instance, reading speed, intelligence, etc). For these nuisance variables, their effect can be neutralized by **randomly assigning** subjects to the different treatment conditions.

Data analysis and conclusion.

In controlled experiments, data are usually analyzed by means of descriptive and inferential statistics. **Descriptive statistics**, such as mean, variance, standard deviation, are designed to describe or summarize a set of data. In order to report significant results and make inference about the population of interest, the descriptive statistics are not sufficient, but some inferential statistic measure is required. Indeed, **inferential statistics** are used to evaluate the statistical hypotheses. These statistics are designed to make inferences about larger populations. The choice of the right statistics to be used depends on the kind of collected data and the questions to be answered.

Parametric statistics are exploited when data are normally distributed. Example of parametric tests are: ANOVA (ANalysis Of VAriance) calculated by

means of F-test or t-test, and linear (or non-linear) regression factor analysis. For instances of the use of F test in adaptive systems see [20, 108], while for examples of t-test see [46, 94, 131].

The *non-parametric statistics* make no assumptions about the distribution of the scores making up a treatment condition. Examples of non-parametric tests are Wilcoxon rank-sum test, rank-sum version of ANOVA, Spearman's rank correlation, Mann-Whitney Test. For examples about the use of non-parametric measures in adaptive systems see [25, 42, 72].

While the above statistics can be applied when the dependent variables to measure are continuous (they can take values as, for instance, time or number of errors, etc), the *Chi square test* (χ^2) instead is the common measure used to evaluate the significant values assumed by categorical data. For example of use of Chi square tests in adaptive systems see [25, 87, 121].

Sensitivity measures should also be calculated. In this context, sensitivity refers to the ability to detect any effects that may exist in the treatments population. The sensitivity of an experiment is given by the effect size and the power. *The effect size or treatment magnitude* (ω^2) measures the strength, or the magnitude, of the treatment effects in the experiment. The *power* of an experiment is the ability to recognize treatment effects. The power can be used for estimating the sample size. Designing the experiments to have a high power rating not only ensures greater repeatability of results, but it makes it more likely to find the desired effects. For an example of sensitivity measures applied to analyze the evaluation results of an adaptive web site see [52], while for details on the importance of sensitivity measures in adaptive and user modeling systems see [32].

Ethnography.

Sustainers of qualitative approaches affirm that laboratory conditions are not real world conditions and that only observing users in natural settings can detect the real behavior of the users. From this perspective, a subject cannot be reduced to a sum of variables and therefore a deeper knowledge of a fewer group of subjects is more useful than an empirical experiment with a representative sample. Qualitative methods of research often make use of ethnographic investigations, also known as participant-observation¹⁴.

Preece et al. [125] classify the ethnographic investigations under the umbrella term "interpretative evaluation". The interpretative evaluation can be best summed up as "spending time with users" and it is based on the assumption that small factors that go behind the visible behavior greatly influence outcomes. According to [125], the interpretative evaluation comes in these flavors:

¹⁴ In social sciences, and in particular in field-study research, participant-observation is a qualitative method of research that requires direct involvement of the researcher with the object of the study. For more details see [140].

- contextual inquiry (see Sec. 1.3.1);
- cooperative evaluation (see Sec. 1.3.2);
- participative evaluation (see Sec. 1.3.2);
- ethnography.

While the first three techniques have been already described, since they should be used in former evaluation phases, ethnography can be better performed in the final evaluation phase.

Ethnography is a qualitative observational technique that is well established in the field of sociology and anthropology. It involves immersing the researcher in the everyday activities of an organization or in the society for a prolonged period of time. Ethnography provides the kind of information that is impossible to gather from the laboratory, since it is concerned with collecting data about real work circumstances. The ethnographic approach in HCI acknowledges the importance of learning more about the way technology is used in real situations [107].

Qualitative methods are seldom applied in the evaluation of adaptive web-based systems. However, statistical analyses are sometimes false, misleading, and too narrow, while insights and qualitative studies do not suffer from these problems as they strictly rely on the users' observed behavior and reactions [111]. Qualitative methods, such as ethnography, could bring fruitful results, especially in order to discover new phenomena (e.g., by observing the users interacting with a web site in their context, new solutions on how to adapt the site can emerge). In fact, qualitative researchers want to comprehend the subjects under study by interpreting their points of view and by analyzing the facts in depth (intensive approach) in order to propose new general understanding of the reality.

Importance for the adaptive web: collection of data in real situations; exploratory studies; discovering new phenomena.

The Grounded Theory.

The Grounded Theory is “a theory derived from data, systematically gathered and analyzed through the research process. In this method, data collection, analysis and eventual theory stand in close relationship to one another. The researcher does not begin a project with a preconceived theory in mind (...). Rather, the researcher begins with an area of study and allows the theory to emerge from the data” [142]. The collected data may be qualitative, quantitative, or a combination of both types, since an interplay between qualitative and quantitative methods is advocated. See Cena, Gena & Modeo [30] for an application of the Grounded Theory methodology with heterogeneous sources of data (both qualitative and quantitative) in an empirical evaluation aimed at choosing a better way to communicate recommendations to the users in the interface for mobile devices. For the development of a cooperative student model in a multimedia application, Grounded Theory has been applied to

understand the many and complex interactions between learners, tutors and learning environment by integrating the range of qualitative and quantitative results collected during the several experimental sessions [8].

Importance for the adaptive web: combined analysis of qualitative and quantitative data; exploratory studies; discovering of new phenomena that can inspire adaptation.

Table 4. Final evaluation phase

	Kind of Factors	Applicability conditions	Pros and cons
Final evaluation phase			
Usability test	Usability of interface adaptation	A running user-adaptive prototype/system	+ objective performance measures and subjective user feedback - adaptive systems require also evaluation of the content layer: usability test is necessary, but not sufficient
Automatic usability testing and web usage mining	Automatic detection of usability of interface adaptation; discovering of sequences of user actions and behavioral patterns	A running user-adaptive system; software for the analysis of log data	+ discovering of real usage of the system; unobtrusive - no information about what the users are thinking/why they do something
Accessibility evaluation	Information on accessibility	A running user-adaptive prototype/system	+ accessibility has always to be tested
Experimental evaluation	Interface (and content) adaptations	A running user-adaptive prototype/system	+ when properly designed gives significant results - possibly artificial lab situation
Ethnography	Collection of data in real work situation	A running user-adaptive system	+ users interact with the system in a real situation - time consuming and expensive
Grounded theory	To combine qualitative and quantitative evaluation, to discover new theories	A running user-adaptive prototype/system	+ comprehensive and explorative - results may be subjective; time consuming

1.4 Key Issues in the Evaluation of Adaptive Systems

Choosing appropriate methods for the evaluation of adaptive web-based systems is crucial. However, when conducting an evaluation study on an adaptive system, a number of issues might arise that are specific for this kind of system. The purpose of this section is to review these issues in order to raise the awareness for the potential problems and to sketch possible counter measures where available. A more in depth discussion of these issues can be found in [154].

1.4.1 Allocation of Resources

Resources required for evaluation studies are frequently underestimated. Setup, data collection and analysis require a high amount of personnel, organizational and sometimes even financial resources [96]. In some cases, small-scale experiments (i.e., assessing every participant for a short time) are not feasible, when adaptation does not happen on the spot, but takes time. The system needs to gather some information about the user before it actually adapts.

However, there are several ways to either reduce the required resources or to assure the allocation of resources in advance. First of all, as described throughout this chapter, it might be useful to spread the evaluation across the whole development cycle. The summative evaluation would then be only a final validation of previous findings under real world conditions. Experience with empirical research has shown that it is a good idea to plan several small experiments or studies rather than a single large one, because this strategy provides more flexibility and limits the risk of flawed experimental designs. Nevertheless, a project proposal should not underestimate the required resources.

Second, several aspects of the evaluation may also be covered by expert assessment rather than user studies. Several of the methods described in this chapter, for instance, cognitive walkthrough (Section 1.3.1) and heuristic evaluation (Section 1.3.2) have been shown to be an effective and efficient way to detect many frequent usability problems with limited resources. There also exist heuristics for the evaluation of adaptivity [93]. However, it should be pointed out that expert evaluations run the risk of being biased if they are conducted by researchers who evaluate their own system.

Third, simulated users might be considered for testing the inference mechanism [71]. If the system is able to distinguish between groups among these simulated users it can at least be assumed to work in the expected way. However, to improve the ecological validity of this kind of study the users should be based on real empirical data.

In the area of information retrieval testing the adaptive system in terms of accuracy, precision and recall with open data sets is a common research method (e.g., [135]). Obviously, simulated users require less resources than real user studies, because the data can be reused in further improvement

cycles and even in the evaluation of other systems. Moreover, the simulation strategy can guarantee that all possible combinations of user characteristics are covered. Therefore, simulated users can be seen as a variant of test cases. However, there are also limitations: simulated users can be used to test the inferences of an adaptive system, but both the user assessment and the effect of the adaptation on the user are excluded. However, if the sample is not based on empirical data, it might deviate from real users in essential aspects. For instance it might contain characteristics or combinations of characteristics that are impossible or that do not exist in the user group.

Finally, cognitive models have been proposed for the evaluation of adaptive systems [96]. A cognitive model is basically a computer program that implements process-oriented specifications of some of the main modules and mechanisms underlying human cognition and social activity [127]. Such a model may interact with an adaptive system and demonstrate important characteristics, e.g., cognitive effort or completion time. The main advantage of this approach is that it facilitates prediction of cognitive processes with variants of the target system without unwanted side effects such as learning, fatigue or reaction. However, adapting a cognitive model to a specific task and environment often requires a lot of effort and expertise even if it is based on an existing cognitive architecture (i.e., a framework for implementing cognitive models).

The last two types of studies (using simulated users and cognitive models) can be categorized as *in silico* experiments [146], a term that has been coined in biology in order to describe experimental settings that are executed in a virtual environment based on computer models (e.g., [161]). Though there are several threats to the validity of *in silico* experiments, they are a powerful and cost-effective strategy if used in combination with *in vivo* (real life) and *in vitro* (laboratory) experiments.

1.4.2 Specification of Control Conditions

Another problem, that is inherent in the evaluation of adaptive systems, occurs when the control conditions of experimental settings are defined. In many studies the adaptive system is compared to a non-adaptive version of the system with the adaptation mechanism switched off [24]. However, adaptation is often an essential feature of these systems and switching the adaptivity off might result in an absurd or useless system [64, 65]. In some systems, in particular if they are based on machine learning algorithms [82, 122, 123], it might even be impossible to switch off the adaptivity.

A preferred strategy might be to compare a set of different adaptation decisions (as far as applicable). Based on the same inferred user characteristics the system can be adapted in different ways. For instance, an adaptive learning system that adapts to the current knowledge of the learner might use a variety of adaptation strategies, including link annotation, link hiding, or curriculum sequencing. Comparing these variants in terms of relevant criteria sketches a

much more complete picture of the adaptation impact than just comparing the standard system with a non-adaptive version. The variants might also include combinations of existing adaptation decisions. However, the variants should be as similar as possible in terms of functionality and layout (often referred to as *ceteris paribus*, all things being equal) in order to be able to trace back the effects to the adaptivity itself. Also matching the context of an experimental setting with real environments seems to be crucial in order to achieve sufficient external validity. Using the example of a recommender system evaluation, Missier & Ricci [39] suggested that it will be necessary to reproduce the real decision environment, i.e., the real system should be tested, with no changes in databases, interface, algorithms, and parameters. Even if this might be a difficult task for some types of adaptation decisions that have an impact on the interaction structure, the interpretability of the results relies a great deal upon these aspects.

1.4.3 Sampling

A proper experimental design requires not only to specify control conditions but also to select adequate samples. On the one hand the sample should be very heterogeneous in order to maximize the effects of the system's adaptivity: the more differences between users, the higher the chances that the system is able to detect these differences and react accordingly. On the other hand, from a statistical point of view, the sample should be very homogeneous in order to minimize the secondary variance and to emphasize the variance of the treatment. It has been reported frequently that too a high variance is a cause of the lack of significance in evaluation studies [21, 96, 104]. For instance, learners in online courses usually differ widely in reading times which might corrupt further comparisons in terms of time savings due to adaptive features. A common strategy to reduce this secondary (undesired) variance is to homogenize or parallelize the sample as much as possible. However, this strategy might be in conflict with the first requirement of sample heterogeneity. The ideal sample would differ widely in terms of the assessed user characteristics but would be homogeneous in terms of all other factors.

A second common strategy to reduce undesired variance is using repeated measurement. The main advantages of this kind of experimental design include: less participants are required, and statistical analysis is based on differences between treatments rather than between groups that are assigned to different treatments. However, this strategy is often not adequate for the evaluation of adaptive systems, because of order effects. If people get used to the first version of the system they might have problems to interact with the second version, because they have built up expectations (a mental model) about the system that are inadequate for the second version. Balancing the order of treatments might alleviate this problem, but the danger of biased results due to unexpected and undesired interactions between the treatments will remain. A third strategy is to control for variables that might have an

impact on the results and to include these variables in the analysis. This strategy, sometimes referred to as dicing, might help to explain results that are diluted by the mean values. E.g., the adaptation decision might be correct for one subgroup, but it has a negative effect for the other subgroup. While the mean value would indicate that there is no effect at all, the detailed analysis demonstrates the strengths and weaknesses of the system. Moreover, there are obviously other criteria that have to be considered when selecting the sample in general. In order to generalize the results the sample should either be representative for the target group or at least not differ from the target group in terms of factors that are known to affect the results (e.g., expertise or motivation). Therefore, samples for evaluation studies with adaptive systems need to be selected carefully.

1.4.4 Definition of Criteria

Current evaluation studies use a broad range of different criteria [153]. The diversity of these criteria inhibits a comparison of different modeling approaches.

The criteria usually taken in consideration for evaluation (e.g., task completion time, number of errors, number of viewed pages) sometimes do not fit the aims of the system. For instance, during an evaluation of a recommender system the relevance of the information provided is more important than the time spent to find it. Another good example is reported by a preliminary study on evaluation of an in-vehicle adaptive system [89]. The results showed that adaptivity is beneficial for routine tasks, while performance of infrequent tasks is impaired. Furthermore, lots of applications are designed for long-time interaction and therefore it is hard to correctly evaluate them in a short and controlled test.

A precise specification of the modeling goals is required in the first place, as this is a prerequisite for the definition of the criteria. The criteria might be derived from the abstract system goals for instance by using the Goal-Question-Metric method (GQM) [148], which systematically defines metrics for a set of quality dimensions in products, processes, and resources. Tobar [144] presented a framework that supports the selection of criteria by separating design perspectives.

Many adaptive web-based systems are concerned with some kind of navigation support. Adaptivity might reduce the complexity of the navigation behavior [75, 155]. Accordingly, accepted graph complexity measures might be used for analyzing the users' behavior. However, as argued by Herder [62], the browsing activity is expected to produce more complex navigation than goal-directed interaction. Therefore, the metrics for the evaluation of user navigation should take into account both the site structure and the kind of user's tasks since, depending on these factors, a reduction in the complexity of the interaction is not necessarily caused by the adaptive behavior of the system. However, as claimed by Krug [83], "It doesn't matter how many times I have to click, as long as each click is a mindless, unambiguous choice". Therefore, if

the web site proposes some kind of adaptation, the adaptive solutions could help the user to disambiguate her choices, reducing the feeling of “being lost in the hyperspace”.

Future research should aim at establishing a set of commonly accepted criteria and assessment methods that can be used independently of the actual user model and inference mechanism in order to explore the strength and weaknesses of the different modeling approaches across populations, domains, and context factors. While current evaluation studies usually yield a single data point in the problem space, common criteria would allow integration of the results of different studies for a broader picture. So-called utility-based evaluation [62] shows how such a comparison across systems could be achieved.

1.4.5 Violation of Accepted Usability Principles

While we argue that the evaluation of adaptive systems must not be seen as being a mere usability testing problem, usability is certainly an important issue. However, several discussions have arisen about the usability of adaptive interfaces [65]. As already sketched in Section 1.3.3 Jameson [73] proposes five usability challenges for adaptive interfaces. These challenges complicate matters for the evaluation of adaptive systems even more, because usability goals and adaptivity goals need to be considered concurrently. For instance, lack of transparency and control can become a threat to the usability of an adaptive system [72, 38]. However, under certain conditions it is possible to match usability and adaptivity goals [74].

1.4.6 Asking for Adaptivity Effects

In many studies the users estimate the effect of adaptivity (e.g., [12]) or rate their satisfaction with the system (e.g., [7, 45, 48] after a certain amount of interaction. However, from a psychological point of view these assessment methods might be inadequate in some situations. Users might have no anchor of what good or bad interaction means for the given task if they do not have any experience with the ‘usual’ non-adaptive way. They might not even have noticed the adaptivity at all, because adaptive action often flows (or should flow) in the subjective expected way rather than in the static predefined way (i.e., rather than prescribing a certain order of tasks or steps, an adaptive system should do what the user wants to do). Therefore, users might notice and report only those events when the system failed to meet their expectations.

On the other hand, qualitative user feedback can be of high value, in particular in early stages of the development. Therefore, surveys and interviews should definitely be considered when planning the assessment, but in order to avoid interpretation problems they should be accompanied by objective measures such as performance, and number of navigation steps. It is highly recommended to at least informally debrief and converse with participants if possible after the trial both from an ethical point of view in order to detect problems such as design.

1.4.7 Separation of Concerns: Layered Evaluation

Comparing an adaptive version with the non-adaptive version in terms of their effectiveness and efficiency might not be a fair test (see Section 1.4.2). Moreover, this design does not provide insights into why the system is better or not.

When designing evaluation studies, it is fundamental to distinguish the different adaptation constituents, and sometimes it might be necessary to evaluate them separately from the beginning. So-called *layered approaches* [22, 76] have been proposed in the literature to separately evaluate the identified adaptation components (layers) of adaptive systems. The cited approaches identify, at least, two layers: the content layer, and the interface layer. This idea comes from Totterdell and Boyle [145], who first phrased the principle of layered evaluation, “*Two types of assessment were made of the user model: an assessment of the accuracy of the model’s inferences about user difficulties; and an assessment of the effectiveness of the changes made at the interface*”. More recent approaches [116, 152, 153] identified several adaptation components and therefore more corresponding evaluation layers, and [115] also proposed specific evaluation techniques to be adopted in every layer. We can see that layered evaluation is one of the peculiarities that characterize the evaluation of adaptive systems, as well as the presence of several *typical users* of the system, to which the system adapts itself. Therefore, groups of significant users should be separately observed across the layers, and the evaluation could underline that adaptive solutions are useful for some users and for others they are not.

1.4.8 Reporting the Results

Even a perfect experimental design will be worthless if the results are not reported in a proper way. In particular statistical data require special care, as the findings might not be interpretable for other researchers if relevant information is skipped. This problem obviously occurs in other disciplines and research areas dealing with empirical findings. Therefore, there are many guidelines and standard procedures for reporting empirical data as suggested or even required by some journals (e.g., [3, 14, 88, 160]). In the special case of adaptive systems, several other things should be reported. First, the inference mechanism should be described in detail, or the reader should at least be referred to a detailed description. Second, the user model should be described in terms of the dimensions or characteristics that are modeled. If applicable the report should contain the theoretically possible values or states of the model as well as the empirically identified states. This is important to characterize both the sample (cf. Section 1.4.3) and the potential impact of the treatment. For instance, if the adaptivity is responsive to user characteristics that occur only once in a while, the impact on the total interaction will be limited. Third, besides statistical standard identifiers (i.e., sample size, means, significance level, confidence interval) the effect size [33] of the treatment is of interest,

because it estimates the adaptivity effect in comparison to the total variance and is therefore an indicator of the utility. It enables practitioners to estimate the expected impact of a new technique or approach and facilitates meta-analyses.

1.5 Conclusions

This chapter has presented a review of methods and techniques for design and evaluation of adaptive web-based systems under a usability engineering perspective. Even though improvement has been registered in a number of evaluation studies in the recent years [51], the evaluation of adaptive web systems needs to reach a more rigorous level in terms of subject sampling, statistical analysis, correctness in procedures, experiment settings, etc. Evaluation studies should benefit from the application of qualitative methods of research and from a rigorous and complete application of user-centered design approach in every development phase of these systems.

To conclude, we advocate the importance of evaluation in every design phase of an adaptive web-based system and at different layers of analysis. Significant testing results can lead to more appropriate and successful systems and the user's point of view can be a very inspiring source of information for adaptation strategies. From our point of view, both quantitative and qualitative methodologies of research can offer fruitful contributions and their correct application has to be carried out by the researchers working in this area in every design phase. Finally, since evaluation in adaptive systems is still in a exploratory phase, new approaches are strongly called for and these can include combining together different techniques, exploring new metrics to assess adaptivity, and adapting the evaluation technique to the adaptive systems features.

References

1. Alepis E., Virvou M., 2006. User Modelling: An Empirical Study for Affect Perception Through Keyboard and Speech in a Bi-modal User Interface. In Proceedings of AH2006, LNCS 4018, pp. 338-341.
2. Alfonseca E. and Rodriguez P., 2003. Modelling Users' Interests and Needs for an Adaptive On-line Information System. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin: Springer, pp. 76-82.
3. Altman D., Gore S., Gardner M., and Pocock S., 1983. Statistical Guidelines for Contributors to Medical Journals. *British Medical Journal*, 286, 1489-1493.
4. Anderson, J. R., Boyle, C. F., and Yost, G., 1985. The Geometry Tutor. 9th International Joint Conference on AI, pp 1-7.

5. Ardissono L., Gena C., Torasso P., Bellifemine F., Chiarotto A., Difino A., Negro B., 2004. User Modeling and Recommendation Techniques for Personalized Electronic Program Guides. In L. Ardissono, A. Kobsa and M. Maybury (Eds.), *Personalization and user-adaptive interaction in digital tv*, Kluwer Academic Publishers, pp. 30-26.
6. Bakeman R. and Gottman J. M., 1986. *Observing Behavior: An Introduction to Sequential Analysis*. Cambridge: Cambridge University.
7. Bares W. H., and Lester J. C., 1997. Cinematographic User Models for Automated Realtime Camera Control in Dynamic 3D Environments. In A. Jameson, C. Paris, and C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97*. Vienna, New York: Springer, pp. 215-226
8. Barker T., Jones S., Britton C., Messer D., 2002. The Use of a Co-operative Student Model of Learner Characteristics to Configure a Multimedia Application. *User Modeling and User-adaptive Interaction* 12(2), pp. 207-241.
9. Baudisch P., Brueckner L., 2002. TV Scout: Lowering the Entry Barrier to Personalized TV Program Recommendation. In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), *Lecture Notes in Computer Science n. 2347: Adaptive Hypermedia and Adaptive Web-Based Systems*. Berlin: Springer, pp. 58-68.
10. Beck J. E., Jia P., Sison J., and Mostow J., 2003. Predicting Student Help-request Behavior in an Intelligent Tutor for Reading. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 303-312.
11. Beck J. E., Jia P., Sison J., and Mostow J., 2003. Assessing Student Proficiency in a Reading Tutor That Listens. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin, etc.: Springer, pp. 323-327.
12. Beck J. E., Stern M., and Woolf B. P., 1997. Using the Student Model to Control Problem Difficulty. In: A. Jameson, C. Paris, and C. Tasso (Eds.), *User modeling: Proceedings of the Sixth International Conference, UM97*. Vienna, New York: Springer, pp. 277-288.
13. Bednarik R., 2005. Potentials of Eye-Movement Tracking in Adaptive Systems. In: Weibelzahl, S. Paramythis, A. and Masthoff Judith (eds.). *Proceedings of the Fourth Workshop on Empirical Evaluation of Adaptive Systems, held at the 10th International Conference on User Modeling UM2005, Edinburgh*, pp. 1-8.
14. Begg C., Cho M., Eastwood S., Horton R., Moher D., Olkin I., Pitkin R., Rennie D., Schultz K., Simel D., and Stroup D., 1996. Improving the Quality of Reporting Randomized Trials (the CONSORT Statement). *Journal of the American Medical Association*, 276(8), 637-639.
15. Bellotti V. and MacLean A. Design Space Analysis (DSA). <http://www.mrc-cbu.cam.ac.uk/amodeus/summaries/DSAsummary.html>
16. Bental D., Cawsey A., Pearson J. and Jones R., 2003. Does Adapted Information Help Patients with Cancer? In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 288-291.
17. Benyon D., 1993. Adaptive Systems: A Solution to Usability Problems. *User Modeling and User-adaptive Interaction* (3), pp. 65-87.
18. Beyer H. and Holtzblatt K., 1998. *Contextual Design: Defining Customer-Centered Systems*, Morgan Kaufmann Publishers, Inc., San Francisco CA.

19. Billsus, D. and Pazzani, M., 1998. Learning Collaborative Information Filters. In: Proceedings of the International Conference on Machine Learning. Morgan Kaufmann Publishers. Madison, Wisc.
20. Brunstein, A., Jacqueline, W., Naumann, A., and Krems, J.F., 2002. Learning Grammar with Adaptive Hypertexts: Reading or Searching? In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), Lecture Notes in Computer Science N. 2347: Adaptive Hypermedia and Adaptive Web-Based Systems. Berlin: Springer.
21. Brusilovsky P., and Eklund J., 1998. A study of user-model based link annotation in educational hypermedia. *Journal of Universal Computer Science, Special Issue on Assessment Issues for Educational Software*, 4(4), 429-448.
22. Brusilovsky P., Karagiannidis C., and Sampson D., 2001. The benefits of layered evaluation of adaptive applications and services. In: S. Weibelzahl, D. N. Chin, & G., Weber (Eds.), *Empirical Evaluation of Adaptive Systems. Proceedings of Workshop At the Eighth International Conference on User Modeling, UM2001*, Pp. 1-8.
23. Brusilovsky P., and Millán, E. 2006. User Models for Adaptive Hypermedia and Adaptive Educational Systems. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, Vol. 4321*. Springer-Verlag, Berlin Heidelberg New York (2006) This Volume.
24. Brusilovsky P., and Pesin L., 1998. Adaptive navigation support in educational hypermedia: An evaluation of the ISIS-tutor. *Journal of Computing and Information Technology*, 6(1), 27-38.
25. Brusilovsky P., Sosnovsky S., Yudelson M., 2006. Addictive Links: The Motivational Value of Adaptive Link Annotation in Educational Hypermedia In Proceedings of AH2006, LNCS 4018, Pp. 51-60.
26. Bull S., 2003. User Modelling and Mobile Learning. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science N. 2702: *User Modeling 2003*. Berlin: Springer, Pp. 383-387.
27. Bunt A., 2005. User Modeling to support user customization. In: Proceedings of UM 2005, LNAI 3538, Pp. 499-501.
28. Burke R., 2002. Hybrid Recommender Systems: Survey and Experiments. *User Modeling and User-Adapted Interaction*. 12(4), Pages 331-370.
29. Burke R., 2006. Hybrid Web Recommender Systems. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, Vol. 4321*. Springer-Verlag, Berlin Heidelberg New York (2006) This Volume.
30. Cena F., Gena C., and Modeo S., 2005. How to communicate recommendations? Evaluation of an adaptive annotation technique. In the Proceedings of the Tenth IFIP TC13 International Conference on Human-Computer Interaction (INTERACT 2005), Pp. 1030-1033.
31. Chapanis, A., 1991. Evaluating usability. In: B. Shackel and S. J. Richardson (Eds.), *Human Factors for Informatics Usability*, Cambridge: Cambridge University, Pp. 359-395.
32. Chin D.N., 2001. Empirical evaluation of user models and user-adapted systems. *User Modeling and User-Adapted Interaction*, 11(1-2), Pp. 181-194.
33. Cohen J., 1977. *Statistical power analysis for the behavioral sciences* (revised ed.). New York: Academic Press.

34. Conati C., Merten C. and Muldner K., 2005. Exploring Eye Tracking to Increase Bandwidth in User Modeling In: Proceedings of UM 2005, LNAI 3538, pp. 357-366.
35. Console L., Gena C. and Torre I., 2003. Evaluation of an On-vehicle Adaptive Tourist Service. In: Weibelzahl, S. and Paramythis, A. (eds.). Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh, pp. 51-60.
36. Courage C., and Baxter K., 2005. Understanding Your Users: A Practical Guide to User Requirements Methods, Tools, and Techniques. Morgan Kaufmann Publishers, San Francisco, CA.
37. Cugini J. Scholtz J., 1999. VISVIP: 3D Visualization of Paths Through Web Sites. In: Proceedings of the International Workshop on Web-Based Information Visualization, pp. 259-263.
38. Czarkowski M., 2005. Evaluating Scrutable Adaptive Hypertext. In: Weibelzahl, S. Paramythis, A. and Masthoff Judith (eds.). Proceedings of the Fourth Workshop on Empirical Evaluation of Adaptive Systems, held at the 10th International Conference on User Modeling UM2005, Edinburgh, pp. 37-46.
39. Del Missier F. and Ricci F., 2003. Understanding Recommender Systems: Experimental Evaluation Challenges. In: Weibelzahl, S. and Paramythis, A. (eds.). Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference, pp. 31-40.
40. Diaper D. (Ed.). Task Analysis for Human-computer Interaction. Chicester, U.K.: Ellis Horwood, 1989.
41. Dimitrova V., 2003. Using Dialogue Games to Maintain Diagnostic Interactions. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin, etc.: Springer, pp. 117-121.
42. Dimitrova V., Self J. and Brna P., 2001. Applying Interactive Open Learner Models to Learning Technical Terminology. In: M. Bauer, P.J. Gmytrasiewicz and J. Vassileva (Eds.), Lecture Notes in Computer Science n. 2109: User Modeling 2001. Berlin, etc.: Springer, pp. 148-157.
43. Dix A., Finlay J., Abowd G. and Beale R., 1998. Human Computer Interaction. Second Edition, Prentice Hall.
44. Dumas J. S. and Redish J. C., 1999. A Practical Guide To Usability Testing. Norwood, N.J. Ablex Publishing Corp.
45. Encarnação L. M., and Stoev S. L., 1999. Application-independent Intelligent User Support System Exploiting Action-sequence Based User Modeling. In: J. Kay (Ed.), User modeling: Proceedings of the Seventh International Conference, UM99. Vienna, New York: Springer, pp. 245-254
46. Farzan R., and Brusilovsky P, 2006 Social Navigation Support in a Course Recommendation System In Proceedings of AH2006, LNCS 4018, pp. 91-100.
47. Fink J., Kobsa A. and Nill A., 1998. Information Provision for All Users, Including Disabled and Elderly People. New Review of Hypermedia and Multimedia, 4, pp. 163-188.
48. Fischer G., and Ye Y., 2001. Personalizing Delivered Information in a Software Reuse Environment. In: M. Bauer, J. Vassileva, and P. Gmytrasiewicz (Eds.), User modeling: Proceedings of the Eighth International Conference, UM2001. Berlin: Springer, pp. 178-187

49. Gena C., 2001. Designing TV Viewer Stereotypes for an Electronic Program Guide. In: M. Bauer, P. J. Gmytrasiewicz and J. Vassileva (Eds.), *Lecture Notes in Computer Science n. 2109: User Modeling 2001*. Berlin, etc.: Springer, pp. 247-276.
50. Gena C., 2003. *Evaluation Methodologies and User Involvement in User Modeling and Adaptive Systems*. Unpub. Phd thesis, Università degli Studi di Torino, Italy.
51. Gena C., 2005. Methods and Techniques for the Evaluation of User-adaptive Systems. *The Knowledge Engineering Review*, Vol 20:1,1-37, 2005.
52. Gena C. and Ardissono L., 2004. Intelligent Support to the Retrieval of Information About Hydric Resources. In: *Proceedings of the International Conference on Adaptive Hypermedia and Adaptive Web-based Systems AH2004*, Eindhoven, The Netherlands, *Lecture Notes in Computer Science*. Pp. 126-135.
53. Gena C., Perna A. and Ravazzi M., 2001. E-tool: a Personalized Prototype for Web Based Applications. In: D. de Waard, K. Brookhuis, J. Moraal and A. Toffetti (Eds.), *Human Factors in Transportation, Communication, Health, and the Workplace*. Shaker Publishing, pp. 485-496.
54. Gena C. and Torre I., 2004. The Importance of Adaptivity to Provide On-Board Services. A Preliminary Evaluation of an Adaptive Tourist Information Service on Board Vehicles. *Special Issue on Mobile A.I. in Applied Artificial Intelligence Journal*.
55. Good N., Schafer J.B., Konstan J.A., Borchers A., Sarwar B.M., Herlocker J.L. and Ricdl J., 1999. Combining Collaborative Filtering with Personal Agents for Better Recommendations. In: *Proceedings of the Sixteenth National Conference on Artificial Intelligence*, pp. 439-446.
56. Gould J. D. and Lewis C., 1983. Designing for Usability – Key Principles and What Designers Think. In: *Human Factors in Computing Systems, CHI '83 Proceedings*, New York: ACM, pp. 50-53.
57. Goy A., Ardissono L. Petrone G., 2006. Personalization in E-Commerce Applications. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*, *Lecture Notes in Computer Science*, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2006) this volume.
58. Greenbaum J. and Kyng M. (Eds.), 1991. *Design At Work: Cooperative Design of Computer Systems*. Lawrence Erlbaum Associates, Inc., Hillsdale, NJ.
59. Greenbaum T. L., 1998. *The Handbook of Focus Group Research (2nd Edition)*. Lexington Books: New York, NY.
60. Habieb-Mammar H., Tarpin-Bernard F. and Prévôt P., 2003. Adaptive Presentation of Multimedia Interface Case Study: Brain Story Course. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 15-24.
61. Hara Y., Tomomune Y., Shigemori M., 2004. Categorization of Japanese TV Viewers Based on Program Genres They Watch. In: L. Ardissono, A. Kobsa and M. Maybury (Eds.), *Personalization and user-adaptive interaction in digital TV*, Kluwer Academic Publishers.
62. Herder E., 2003. Utility-Based Evaluation of Adaptive Systems. In: *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems*,

- held at the 9th International Conference on User Modeling UM2003, Pittsburgh, pp. 25-30.
63. Herder E., Weinreich H., Obendorf H., Mayer M., 2006. Much to Know About History. In Proceedings of AH2006, LNCS 4018, pp. 283-287.
 64. Höök K., 1997. Evaluating the Utility and Usability of an Adaptive Hypermedia System. In: Proceedings of 1997 International Conference on Intelligent User Interfaces, ACM, Orlando, Florida, 179-186.
 65. Höök K., 2000. Steps to Take Before IUIs Become Real. *Journal of Interacting with Computers*, 12(4), 409-426.
 66. Iglezakis D., 2005. Is the ACT-value a Valid Estimate for Knowledge? An Empirical Evaluation of the Inference Mechanism of an Adaptive Help System. In: Weibelzahl, S. Paramythis, A. and Masthoff J. (eds.). Proceedings of the Fourth Workshop on Empirical Evaluation of Adaptive Systems, held at the 10th International Conference on User Modeling UM2005, Edinburgh, pp. 19-26.
 67. Ikehara S., Chin D.N. and Crosby M. E., 2003. A Model for Integrating an Adaptive Information Filter Utilizing Biosensor Data to Assess Cognitive Load. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 208-212.
 68. International Organisation for Standardisation ISO 9241-11:1998 Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs), Part 11 Guidance on Usability, 1998. <http://www.iso.org/>
 69. International Organisation for Standardisation ISO/TS 16071:2003 Ergonomics of Human-system Interaction - Guidance on Software Accessibility. Technical Specification, 2003. <http://www.iso.org/>
 70. Jackson T., Mathews E., Lin D., Olney A. and Graesser A., 2003. Modeling Student Performance to Enhance the Pedagogy of AutoTutor. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 368-372.
 71. Jameson A., Schäfer R., Weis T., Berthold A., and Weyrath T., 1999. Making Systems Sensitive to the User's Changing Resource Limitations. *Knowledge-Based Systems*, 12(8), 413-425.
 72. Jameson A., Schwarzkopf E., 2003. Pros and Cons of Controllability: An Empirical Study. In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), *Proceedings of AH'2002, Adaptive Hypermedia and Adaptive Web-Based Systems*, Springer, 193-202.
 73. Jameson A., 2003. Adaptive Interfaces and Agents. *The Human-Computer Interaction Handbook*, Lawrence Erlbaum Associates, New Jersey, pp. 316-318.
 74. Jameson A., 2005. User Modeling Meets Usability Goals. In: Proceedings of UM 2005, LNAI 3538, pp. 1-3.
 75. Juvina I. and Herder E., 2005. The Impact of Link Suggestions on User Navigation and User Perception. In: Proceedings of UM 2005, LNAI 3538, pp. 483-492.
 76. Karagiannidis C. and Sampson D., 2000. Layered Evaluation of Adaptive Applications and Services. In: P. Brusilovsky, O. Stock, C. Strapparava (Eds.), *Adaptive Hypermedia and Adaptive Web-Based Systems, Lecture Notes in Computer Science Vol.1892*, pp. 343-346.
 77. Katz-Haas R., 1998. Ten Guidelines for User-Centered Web Design, *Usability Interface*, 5(1), pp. 1213.

78. Kay J., 2001. Learner Control. *User Modeling and User-Adapted Interaction*, Tenth Anniversary Special Issue, 11(1-2), Kluwer, 111-127.
79. Keppel G., 1991. *Design and Analysis: A Researcher's Handbook*. Englewood Cliffs, NJ: Prentice-Hall.
80. Keppel G., Sauffley W. H. and Tokunaga H., 1998. *Introduction to Design and Analysis, A Student's Handbook*. Second Edition, Englewood Cliffs, NJ: Prentice-Hall.
81. Kirby M.A.R., 1991. *CUSTOM Manual Dpo/std/1.0*. Huddersfield: HCI Research Centre, University of Huddersfield.
82. Krogsæter M., Oppermann R. and Thomas C. G., 1994. A User Interface Integrating Adaptability and Adaptivity. In: R. Oppermann (Ed.), *Adaptive user support*. Hillsdale: Lawrence Erlbaum, pp. 97-125
83. Krug S., 2000. *Don't Make Me Think! A Common Sense Approach to Web Usability*. Indianapolis, IN: New Riders Publishing.
84. Krüger A., Heckmann D., Kruppa M., Wasinger R., 2006. Web-based Mobile Guides. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*, Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2006) this volume.
85. Kuffik T., Shapira B., Elovici Y. and Maschiach A., 2003. Privacy Preservation Improvement By Learning Optimal Profile Generation Rate. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science n. 2702: User Modeling 2003*. Berlin: Springer, pp. 168-177.
86. Kumar A., 2006. Evaluating Adaptive Generation of Problems in Programming Tutors Two Studies. In the Proc. of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, 2006, pp. 450-459.
87. Kurhila J., Miettinen M., Nokelainen P., Tirri H., 2002. EDUCO - A Collaborative Learning Environment Based on Social Navigation. In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), *Lecture Notes in Computer Science n. 2347: Adaptive Hypermedia and Adaptive Web-Based Systems*. Berlin: Springer, pp. 242-252.
88. Lang T. and Secic M., 1997. *How to Report Statistics in Medicine: Annotated Guidelines for Authors, Editors and Reviewers*. Philadelphia, PA: American College of Physicians.
89. Lavie T., J. Meyer, K. Bengler, and J. F. Coughlin, 2005. The Evaluation of In-Vehicle Adaptive Systems. In: Weibelzahl, S. Paramythis, A. and Masthoff J. (eds.). *Proceedings of the Fourth Workshop on Empirical Evaluation of Adaptive Systems*, held at the 10th International Conference on User Modeling UM2005, Edinburgh, pp. 9-18.
90. Lee J. and Lai K.-Y., 1991. What's in Design Rationale? *Human-Computer Interaction special issue on design rationale*, 6(3-4), pp. 251-280.
91. Litman D. J., and Pan, S., 2000. Predicting and Adapting to Poor Speech Recognition in a Spoken Dialogue System. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence* (pp. 722-728). Austin, TX.
92. Magnini B. and Strapparava C., 2001. Improving User Modelling with Content-based Techniques. In: *UM2001 User Modeling: Proceedings of 8th International Conference on User Modeling (UM2001)*, Sonthofen (Germany), July 2001. Springer Verlag.
93. Magoulas G. D., Chen S. Y. and Papanikolaou K. A., 2003. Integrating Layered and Heuristic Evaluation for Adaptive Learning Environments. In:

- Weibelzahl, S. and Paramythis, A. (eds.). Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems, held at the 9th International Conference on User Modeling UM2003, Pittsburgh, pp. 5-14.
94. Martin B. and Mitrovic T., 2002. WETAS: A Web-Based Authoring System for Constraint-Based ITS . In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), Lecture Notes in Computer Science n. 2347: Adaptive Hypermedia and Adaptive Web-Based Systems. Berlin: Springer, pp. 396-305.
 95. Martin B., Mitrovic T., 2006. The Effect of Adapting Feedback Generality in ITS E-Learning and Personalization In Proceedings of AH2006, LNCS 4018, pp. 192-202.
 96. Masthoff, J. (2002). The Evaluation of Adaptive Systems. In N. V. Patel (Ed.), Adaptive evolutionary information systems. Idea Group publishing. pp329-347.
 97. Masthoff, J. (2006). The User As Wizard. In the Proc. of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, 2006, pp. 460-469.
 98. Matsuo Y., 2003. Word Weighting Based on User's Browsing History. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin, etc.: Springer, pp. 35-44.
 99. Maulsby, D., Greenberg, S. and Mander, R. (1993) Prototyping an Intelligent Agent Through Wizard of Oz. In ACM SIGCHI Conference on Human Factors in Computing Systems, Amsterdam, The Netherlands, May, p277-284, ACM Press.
 100. Maybury M. and Brusilovsky P. (Eds.), 2002. The Adaptive Web, Volume 45. Communications of the ACM.
 101. McCreath E. and Kay J., 2003. IEMS : Helping Users Manage Email. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin: Springer, pp. 263-272.
 102. McNee S. M., Lam S. K, Konstan J. A. and Riedl J., 2003. Interfaces for Eliciting New User Preferences in Recommender Systems. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin: Springer, pp. 178-187
 103. Mitrovic A., 2001. Investigating Students' Self-assessment Skills. In: M. Bauer, P. J. Gmytrasiewicz and J. Vassileva (Eds.), Lecture Notes in Computer Science n. 2109: User Modeling 2001. Berlin: Springer, pp. 247-250.
 104. Mitrovic A. and Martin B. 2002. Evaluating the Effects of Open Student Models on Learning. In: P. DeBra, P. Brusilovsky and R. Conejo (eds), Proceedings of Adaptive Hypermedia and Adaptive Web-Based Systems, Berlin: Springer, pp. 296-305.
 105. Mobasher R., 2006. Data Mining for Web Personalisation. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): The Adaptive Web: Methods and Strategies of Web Personalization, Lecture Notes in Computer Science, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2006) this volume.
 106. Mobasher R., Cooley R., and Srivastava J., 2000. Automatic Personalization Based on Web Usage Mining. Communications of the ACM, (43) 8, 2000.
 107. Monk A., Wright P., Haber J. and Davenport L., 1993. Improving Your Human Computer Interface: A Practical Approach. BCS Practitioner Series, Prentice-Hall International, Hemel Hempstead.
 108. Müller C., Großmann-Hutter B., Jameson A., Rummer R., Wittig F., 2001. Recognizing Time Pressure and Cognitive Load on the Basis of Speech: An

- Experimental Study. In: M. Bauer, P. J. Gmytrasiewicz and J. Vassileva (Eds.), *Lecture Notes in Computer Science* n. 2109: User Modeling 2001. Berlin, etc.: Springer, pp 24-33.
109. Nielsen J., 1993. *Usability Engineering*. Boston, MA, Academic Press.
 110. Nielsen J., 2000. Why You Only Need to Test With 5 Users. In: *Alertbox*, <http://www.useit.com/alertbox/20000319.html>.
 111. Nielsen J., 2004. Risks of Quantitative Studies. In: *Alertbox*, <http://www.useit.com/alertbox/20040301.html>.
 112. Nielsen J. and Mack R. L. (Eds.), 1994. *Usability Inspection Methods*. New York, NY: John Wiley & Sons.
 113. Nielsen J. and Molich R., 1990. Heuristic Evaluation of User Interfaces. In: *Proceedings of CHI '90*, Seattle, Washington, pp. 249-256.
 114. Norman D.A. and Draper S.W., 1986. *User Centered System Design: New Perspective on HCI*. Hillsdale NJ, Lawrence Erlbaum.
 115. Paramythis A., Totter A. and Stephanidis C., 2001. A Modular Approach to the Evaluation of Adaptive User Interfaces. In: S. Weibelzahl, D. Chin, and G. Weber (Eds.). *Proceedings of the First Workshop on Empirical Evaluation of Adaptive Systems*, Sonthofen, Germany, pp. 9-24.
 116. Paramythis A., and Weibelzahl S., 2005. A Decomposition Model for the Layered Evaluation of Interactive Adaptive Systems. In: *Proceedings of UM 2005*, LNAI 3538, pp. 438-442.
 117. Paganelli L. and Paterno' F., 2002. Intelligent Analysis of User Interactions with Web Applications. In: *Proceedings of the 2002 International Conference on Intelligent User Interfaces*. ACM Press.
 118. Pazzani M. J. and Billsus D., 2006. Content-based Recommendation Systems. In Brusilovsky, P., Kobsa, A., Nejdl, W. (eds.): *The Adaptive Web: Methods and Strategies of Web Personalization*, *Lecture Notes in Computer Science*, Vol. 4321. Springer-Verlag, Berlin Heidelberg New York (2006) this volume.
 119. Perkowski M., and Etzioni O., 2000. Adaptive Web Sites. *Communications of the ACM*, 43(8), 2000, pp. 152-158.
 120. Pierrakos D., Paliouras G., Papatheodorou, C., and Spyropoulos, C.D., 2003. Web Usage Mining As a Tool for Personalization: a Survey. *International Journal of User Modeling and User-Adapted Interaction*, 13(4), pp. 311-372.
 121. Pirolli P. and Fu W. T., 2003. SNIF-ACT: A Model of Information Foraging on the World Wide Web. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), *Lecture Notes in Computer Science* n. 2702: User Modeling 2003. Berlin: Springer, pp. 45-54.
 122. Pohl, W., 1997. LaboUr-machine Learning for User Modeling. In: M. J. Smith, G. Salvendy, and R. J. Koubek (Eds.). *Design of computing systems: Social and ergonomic considerations. proceedings of the seventh international conference on human-computer interaction*. Amsterdam: Elsevier, Vol. B, pp. 27-30
 123. Pohl, W., 1998. User-adapted Interaction, User Modeling, and Machine Learning. In: U. J. Timm and M. Rössel (Eds.), *Proceedings of the sixth german workshop on adaptivity and user modeling in interactive systems*, ABIS98. Erlangen.
 124. Polson P.G., Lewis C., Rieman J. and Wharton C., 1992. Cognitive Walk-throughs: A Method for Theory- Based Evaluation of User Interfaces. *International Journal of Man-Machine Studies* 36, 741-773.

125. Preece J., Rogers Y., Sharp H. and Benyon D., 1994. Human-computer Interaction. Addison-Wesley Pub.
126. Resnick P. and Varian H. R., 1997. Special Issue on Recommender Systems. Communications of the ACM, 40, 1997.
127. Ritter F. E., Shadbolt N., Elliman D., Young R., Gobet F. and Baxter G., 2002. Techniques for Modeling Human and Organizational Behaviour in Synthetic Environments: A Supplementary Review. Wright-Patterson Air Force Base, OH: Human Systems Information Analysis Center.
128. Rizzo P., Lee I.H., Shaw E., Johnson W.L., Wang N. and Mayer R.E., 2005. A Semi-Automated Wizard of Oz Interface for Modeling Tutorial Strategies In: Proceedings of UM 2005, LNAI 3538, pp. 174-178.
129. Rosenfeld L. and Morville P., 1998. Information Architecture. O'Reilly.
130. Rubin J., 1994. Handbook of Usability Testing: How to Plan, Design, and Conduct Effective Tests. John Wiley & Sons; 1 edition.
131. Santos Jr. E., Nguyen H., Zhao Q. and Pukinskis E., 2003. Empirical Evaluation of Adaptive User Modeling in a Medical Information Retrieval Application. In: P. Brusilovsky, A. Corbett and F. De Rosis (Eds.), Lecture Notes in Computer Science n. 2702: User Modeling 2003. Berlin: Springer, pp. 292-296.
132. Sarwar B. M., Konstan J. A., Borchers A., Herlocker J., Miller B. and Riedl J., 1998. Using Filtering Agents to Improve Prediction Quality in the GroupLens Research Collaborative Filtering System. In: Proceeding of the ACM Conference on Computer Supported Cooperative Work (CSCW), Seattle, WA, pp. 439-446.
133. Scarano V., Barra M., Maglio P. and Negro A., 2002. GAS: Group Adaptive Systems. In: P. De Bra, P. Brusilovsky and R. Conejo (Eds.), Lecture Notes in Computer Science n. 2347: Adaptive Hypermedia and Adaptive Web-Based Systems. Berlin: Springer, pp. 47-57.
134. Sears A. and Shneiderman B., 1994. Split Menus: Effectively Using Selection Frequency to Organize Menus. ACM Transactions on Computer-Human Interaction, 1, 1, pp. 27-51.
135. Semeraro G., Ferilli S., Fanizzi N. and Abbattista F., 2001. Learning Interaction Models in a Digital Library Service. In: M. Bauer, J. Vassileva, and P. Gmytrasiewicz (Eds.), User modeling: Proceedings of the Eighth International Conference, UM2001. Berlin: Springer, pp. 44-53.
136. Shardanand U. and Maes P., 1995. Social Information Filtering for Automating "Word of Mouth". In: Proceedings of CHI-95, Denver, CO, pp. 210-217.
137. Smyth B. and Cotter P., 2002. Personalized Adaptive Navigation for Mobile Portals. In: Proceedings of the 15th European Conference on Artificial Intelligence - Prestigious Applications of Intelligent Systems, Lyons, France, 2002.
138. Specht M. and Kobsa A., 1999. Interaction of Domain Expertise and Interface Design. In: Adaptive Educational Hypermedia, Workshops on Adaptive Systems and User Modeling on the World Wide Web at WWW-8, Toronto, Canada, and Banff, Canada.
139. Spiliopoulou M., 2000. Web Usage Mining for Web Site Evaluation. Communications of the ACM, (43) 8, 2000.
140. Spradley J., 1980. Participant Observation. Wadsworth Publishing.
141. Stephanidis C., 2001. Adaptive Techniques for Universal Access. User Modeling and User-Adapted Interaction, Volume 11, Issue 1-2, pp. 159 - 179.

142. Strauss A. L. and Corbin J. M., 1998. *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*. SAGE, Thousand Oaks.
143. Tasso C. and Omero P., 2002. *La Personalizzazione Dei Contenuti Web*. Franco Angeli, Milano, Italy.
144. Tobar C. M., 2003. Yet Another Evaluation Framework. In: S. Weibelzahl and A. Paramythis (Eds.), *Proceedings of the Second Workshop on Empirical Evaluation of Adaptive Systems*, held at the 9th International Conference on User Modeling UM2003. Pittsburgh, pp. 15-24
145. Totterdell P., and Boyle E., 1990. The Evaluation of Adaptive Systems. In: D. Browne, P. Totterdell and M. Norman (Eds.), *Adaptive User Interfaces*. London: Academic Press, pp. 161-194.
146. Travassos G. H., and Barros M. O., 2003. Contributions of In Virtuo and In Silico Experiments for the Future of Empirical Studies in Software Engineering. In: A. Jedlitschka and M. Ciolkowski (Eds.), *Proceedings of the Second Workshop on the Future of Empirical Studies in Software Engineering* (pp. 109-121). Roman Castles, Italy.
147. Uruchurtu E., MacKinnon L., Rist R., 2005. User Cognitive Style and Interface Design for Personal, Adaptive Learning In: *Proceedings of UM2005*, LNAI 3538, pp. 154-163.
148. van Solingen R., and Berghout E., 1999. *The Goal/question/metric Method: A Practical Guide for Quality Improvement of Software Development*. London: McGraw-Hill.
149. van Barneveld, J. and van Setten, M. (2003). Involving users in the design of user interfaces for TV recommender systems. 3rd Workshop on Personalization in Future TV, Associated with UM03, Johnstown, PA
150. Web Accessibility Initiative W3C - Evaluating Web Sites for Accessibility [Http://www.w3.org/WAI/eval/](http://www.w3.org/WAI/eval/)
151. Web Accessibility Initiative W3C - Web Content Accessibility Guidelines 1.0 [Http://www.w3c.org/TR/WAI-WEBCONTENT](http://www.w3c.org/TR/WAI-WEBCONTENT)
152. Weibelzahl S., 2001. Evaluation of Adaptive Systems. In: M. Bauer, P. J. Gmytrasiewicz and J. Vassileva (Eds.), *Lecture Notes in Computer Science N. 2109: User Modeling 2001*. Berlin: Springer, Pp. 292-294.
153. Weibelzahl S., 2003. *Evaluation of Adaptive Systems*. Dissertation. University of Trier, Germany.
154. Weibelzahl S., 2005. Problems and pitfalls in the evaluation of adaptive systems. In: S. Chen and G. Magoulas (Eds.), *Adaptable and Adaptive Hypermedia Systems* (pp. 285-299). Hershey, PA: IRM Press
155. Weibelzahl S. and Lauer C.U., 2001. Framework for the Evaluation of Adaptive CBRSystems. In: U. Reimer, S. Schmitt, and I. Vollrath (Eds.), *Proceedings of the 9th German Workshop on Case-Based Reasoning (GWCBR01)*, Aachen: Shaker, Pp. 254-263.
156. Weibelzahl S. and Weber G., 2001. A database of empirical evaluations of adaptive systems. In: R. Klinkenberg, S. Rping, A. Fick, N. Henze, C. Herzog, R. Molitor, and O. Schrder (Eds.), *Proceedings of Workshop Lernen - Lehren - Wissen - Adaptivitt (LLWA 01)*, Research Report in Computer Science Nr. 763, University of Dortmund, Pp. 302-306.
157. Weibelzahl S., Jedlitschka2 A., and Ayari B., 2006. Eliciting Requirements for a Adaptive Decision Support System through Structured User Interviews.

- In the Proc. of the Fifth Workshop on User-Centred Design and Evaluation of Adaptive Systems, 2006, Pp. 470-478.
158. Weibelzahl S. and Weber, G., 2003. Evaluating the inference mechanism of adaptive learning systems. In: Lecture Notes in Computer Science N. 2702: User Modeling 2003, Berlin: Springer, Pp. 154-162.
 159. Whiteside J., Bennett J., and Holtzblatt K., 1988. Usability Engineering: Our Experience and Evolution. In: M. Helander (ed.). Handbook of Human-Computer Interaction, New York: North-Holland, 1988, Pp. 791-817.
 160. Wilkinson, L., and Task Force on Statistical Inference, 1999. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist*, 54(8), 594-604.
 161. Wingender, E. (Ed.), 1998. *In Silico Biology*. An international journal on computational molecular biology. IOS Press.