Lex van Velsen

# User-Centered Design for Personalization

# User-Centered Design for Personalization

**Lex S. van Velsen**

# USER-CENTERED DESIGN FOR PERSONALIZATION

## PROEFSCHRIFT

ter verkrijging van
de graad van doctor aan de Universiteit Twente,
op gezag van de rector magnificus,
prof. dr. H. Brinksma,
volgens besluit van het College voor Promoties
in het openbaar te verdedigen
op 25 februari 2011 om 14.45 uur

Lex Stefan van Velsen
geboren op 23 maart 1982
te Nijmegen

Samenstelling Promotiecommissie

| | |
|---|---|
| Promotor | Prof. Dr. M.F. Steehouder |
| Assistent-promotor | Dr. T.M. van der Geest |
| Leden | Prof. Dr. P.A.E. Brey |
| | Prof. Dr. J.A.G.M. van Dijk |
| | Prof. Dr. E.J. Krahmer |
| | Prof. Dr. M.A. Neerincx |
| | Dr. A. Paramythis |

# Contents

# Chapter 1

## Introduction

*"Any sufficiently advanced technology is indistinguishable from magic"*
-- Arthur C. Clarke

## 1.1 Introduction

It is the year 2054 and John Anderton enters the subway station. A camera films his entrance and a central computer recognizes him as John Anderton. As a result, large screens show commercials that address him personally. The first one is for a car: "A road diverges in the desert. Lexus. The road you're on, John Anderton, is the one less-traveled. Make sure you…" John is out of earshot before he can hear the end of it. "John Anderton, You could use a Guinness right about now!" is shouted at him while a screen on his left shows five huge glasses of the Irish stout. In this classic scene of the 2002 movie 'Minority report' by Steven Spielberg, the viewer is treated to a glimpse of the future that contains personalized advertisements.

Although the movie takes place in the year 2054 and can be classified as science fiction, personalized advertisements are by no means a future scenario only. They can already be found in the form of personal recommendations provided by online stores like Amazon and Netflix. But nowadays, personalization can also be found in other formats. Governments provide their citizens with personalized portals and museums provide their guests with personalized tours which can be consulted on handheld devices, to name just a few examples.

The essence of personalization is that communication is geared towards an individual's characteristics, preferences and context. In the current communication landscape this is often done electronically. This heavy focus on the individual has its consequences for design. How can the correspondence between electronically personalized communication and the individual be optimized? How does one deal with delicate issues like privacy, trust and the need for control? And how do you evaluate a website when it looks different to each individual?

This thesis focuses on attuning the user-centered design approach to the context of electronic personalization. Four studies will show how design and evaluation methods can be brought into action during the different phases of the user-centered design process of electronic personalization and can tackle the implications of dealing with electronic communication that is tailored to the individual. In this chapter, we will introduce the reader to the main concepts of this thesis and their origins: personalization and user-centered design.

## 1.2 Personalization: An overview

The idea of personalizing electronic communication arose in the early 1980s (Weibelzahl, 2003). According to Brusilovsky (2001), the first research on personalization[1] dates to the early 1990s, with the amount of research done on the topic taking off after 1996. This was due to the growing popularity of the World Wide Web and the possibilities it offered for creating personalized media content. Furthermore, by then researchers realized that personalization proved to add value and was therefore worth pursuing. Finally, around this time, the commercial sector realized that electronic personalization could be a fruitful replacement of the mass marketing techniques applied up to that point. Hence, the use of personalized marketing features was introduced, thereby offering personalization to the public at large (Kobsa, 2001).

Although personalization can have different goals and can make use of different instruments, its basic workings are roughly the same. We will now elaborate on the two phases that are elemental in the process of creating tailored communication: user modeling and personalizing output.

### 1.2.1 User modeling

Before system output can be personalized, for each user a file must be created, called a user model. In this model, information about a particular user is stored. On the basis of the information stored in the user model, the system determines if output needs to be tailored for the individual and, if so, in what form. It is also possible to tailor output to a homogeneous group of users. In this case, the personalization of output is based upon a group model: a file containing information about a particular group of users.

User modeling is concerned with the creation of a valid model of an individual user. Based on Kobsa, Koenemann and Pohl (2001), we list the kinds of data that can be used to create a user model:

1. User data:
   § Demographic data
   § User knowledge
   § User skills and capabilities
   § User interests and preferences
   § User goals and plans
2. Usage data:

---

[1] Whenever I talk of personalization in this thesis, I mean personalization done by interactive systems, unless explicitly stated otherwise.

§ User clicking
§ User viewing times
§ User ratings
§ User tags
§ User purchases or related actions
§ Browser actions (e.g., saving, printing)
3. Environment data:
§ Software environment
§ Hardware environment
§ User location

These data can be collected implicitly and/or explicitly. If data are collected only implicitly, they are inferred from user behavior. When personalization is based upon implicitly collected user data, the system is *adaptive*. Users can also explicitly state what they would like the personal output to look like, which is then stored in the user model. In this case, a system is *adaptable*. Many personalized systems offer adaptive as well as adaptable features (Wu, Im, Tremaine, Instone, & Turoff, 2003).

A personalized system collects one or more kinds of data and then applies rules to interpret these kinds of data and to make inferences based on this data. For example, if John uses an online bookstore to purchase biographies of the painters Van Gogh, Monet, and Renoir, the system may deduce that John is interested in books about Impressionist painters. Consequently, this inference is stored in John's user model. To discuss the methods of acquiring and interpreting the kinds of data listed above would be a technical matter and outside the scope of this thesis. We refer those who are interested to Kobsa et al. (2001).

### 1.2.2 Personalizing output

Once a user model is created, it can be used to decide whether or not to tailor output. If the rules in a system lead to the decision to tailor output for an individual, many different techniques can be used. Several overviews of these techniques have been published (Brusilovsky, 1996, 2001; Knutov, De Bra, & Pechenizkiy, 2009; Kobsa, Koenemann, & Pohl, 2001) that display a large degree of overlap. Based on these overviews, we list the possible forms of personalized output.

**Adaptation of content**. This type of personalization deals with tailoring the content of an entire or parts of a communication message (e.g., a Web page or a video), or one or more fragments thereof. In the first case, there will be different messages prepared for different kinds of users, and the system will decide which message will be presented to each user. When one or more

fragments of the message will be personalized, there exists a general message that will be presented to all users, but certain parts will be tailored by, for example, leaving out parts or rearranging the text in the message to better suit the receiver.

Examples: Amazon's book recommendations; the adaptable homepages of major search engines like Google (*i*Google) and Yahoo! (My Yahoo!).

**Adaptation of presentation**. This type of personalization deals with tailoring the layout of a message or the modality in which it is presented.

Examples: A Web site that provides content in different modalities to print-disabled users; a Web site that only shows text when accessed by means of a mobile phone.

**Adaptation of navigation**. This type of personalization deals with tailoring the way in which a user navigates through a system (e.g., a Web site) or through the Internet in general. In the case of a closed hyperspace like a Web site, the adaptation can take the form of creating personalized tours, hiding links, or sorting links personally. Personalizing navigation in an open hyperspace, like the World Wide Web, is mostly done by means of personalized search engines.

Examples: A search engine that removes results that are irrelevant for a specific user; a digital museum guide that only displays art pieces of the user's favorite artists.

**Adaptation of user input**. This type of personalization deals with tailoring the text in entry fields, which originally had to be filled in by users themselves. This text can either be incorporated from a user's user model or be collected from a connected system in which the user also has a user model and the required information is already known. Furthermore, information submitted by the user can be expanded with user-related data.

Examples: Pre-filled online government forms; automated tagging of photos uploaded to a photo sharing service.

### 1.2.3 A definition of personalization

Based on our discussion of user modeling and personalizing output, we define *personalized systems* by expanding on the definition of an adaptive system given by Benyon & Murray (1993).

Personalized systems are systems that can alter aspects of their content, structure, functionality or interface on the basis of a user model generated from implicit and/or explicit user input, in order to accommodate the differing needs of individuals or groups of users and the changing needs of users over time.

In this section, we have described the generation of personalized system output, a process that requires several steps, such as user modeling and personalizing output. This makes it different from the generation of "traditional" one-size-fits-all output, which is relatively straightforward. Personalization can be seen as a specific way of analyzing the audience and, consequently, tailoring communication. In that sense, personalization is not only a technical process, but also a rhetorical process.

## 1.3 Personalization, rhetoric, and the audience

In order to get to the source of personalization, we must go back to ancient Greece. In Phaedrus, which Peters (1999) characterizes as the first book on communication science, Socrates and Phaedrus discuss love and the foundations of rhetoric (Plato, trans. 2005). While discussing these foundations, a fictive Socrates states:

"Since the power of speech is in fact a leading of the soul, the man who means to be an expert in rhetoric must know how many forms soul has. Thus their number is so and so, and they are of such and such kinds, which is why some people are like this, and others like that; and these having been distinguished in this way, then again there are so many forms of speeches, each one of such and such a kind. People of one kind are easily persuaded for one sort of reason by one kind of speech to hold one kind of opinion, while people of another kind are for some others sorts of reasons difficult to persuade" (Plato, trans. 2005, p. 271, c10–d5).

Socrates explains here that people are not alike, but are individuals with unique characteristics, or small groups of similar individuals. Each individual or small homogeneous group is best persuaded by applying a tailored rhetorical approach.

After stating that there are different kinds of people who require different kinds of persuasion, Socrates describes the competences a rhetorician needs to create a speech that is tailored to the characteristics of the listener and that thereby achieves successful persuasion.

"…when he both has sufficient ability to say what sort of man is persuaded by what sorts of things, and is capable of telling himself when he observes him that this is the man, this the nature of person that was discussed before, now actually present in front of him, to whom he must now apply these

kinds of speech in this way in order to persuade him of this kind of thing when he now has all of this, and has also grasped the occasions for speaking and for holding back, and again for speaking concisely and piteously and in an exaggerated fashion, and for all the forms of speeches he may learn, recognizing the right and the wrong time for these, then his grasp of the science will be well and completely finished, but not before that" (Plato, trans. 2005, p. 271, e1–272, a5).

The competences that Socrates mentions also describe the steps by which a rhetorician must tailor a speech. First, the rhetorician has to identify the individual listener ("this is the man"). The rhetorician then needs to get to know and understand this individual listener ("this [is] the nature of person […] now actually present in front of him"). For each individual listener, the rhetorician can decide upon a suitable goal to be achieved by means of rhetoric ("to persuade him of this kind of thing"). Taking the individual listener's characteristics and the goal to be achieved into consideration, the rhetorician needs to decide upon a suitable communication strategy ("he must now apply these kinds of speech"). And even these strategies can be tailored into specific presentation forms ("apply these kinds of speech in this way in order to persuade him"). In short, the steps to create a personalized message are, according to Socrates:
1. Identify the individual.
2. Get to know the individual.
3. Set a communication goal for the individual.
4. Tailor the rhetorical approach to the individual.
5. Tailor the communication content to the individual.
Interestingly, these steps resemble the steps in the personalization process as performed by many personalized systems. In Table 1.1, we have listed the rhetorical steps to personalization side by side with the steps of the technical personalization process, as characterized in Paramythis and Weibelzahl (2005). The table shows that in both approaches to personalization, first, the user is identified. Then, the rhetorician has to get to know him or her, or a user model has to be created. Next, a communication goal is set, while in the technical counterpart it is decided whether personalization is appropriate in a given situation and what this personalization should entail. And finally, the actual content of the message is tailored.

Although the steps in both processes are very similar, the means by which the personalized message is conveyed are very different. Socrates argued that tailoring a speech to the individual can only be done by means of personal conversations (Peters, 1999). The written word, or broadcasting

in general, is to be considered an inferior means of communication, as the message to be communicated cannot be geared to the characteristics of an individual, and thereby loses persuasive strength.

"And when once it is written, every composition trundles about everywhere in the same way, in the presence both of those who know about the subject and of those who have nothing at all to do with it, and it does not know how to address those it should address and not those it should not" (Plato, trans. 2005, p. 275, e1).

Table 1.1 A comparison of rhetorical steps and the personalization process

| Rhetorical Steps | Personalization Process |
| --- | --- |
| Identify the individual | Identify user |
| Get to know the individual | Collect user data |
| | Interpret user data |
| Set a communication goal for the individual | Decide upon personalization |
| Tailor the rhetorical approach to the individual Tailor the communication content to the individual | Apply adaptation |

Socrates believed personalized messages to be more persuasive than general ones. And for many centuries, face-to-face communication was the only means to guarantee that personalization could be successful. However, the possibilities for tailoring mediated messages to an audience (or to audience segments) have changed due to the evolving nature of audiences, new methods of analyzing these audiences, and advances in technology. Ultimately, this has led to a situation in which personalization can be achieved electronically. In the next sections, we will set out how the view on "the audience" has evolved. This will show how the ancient starting point (personalization by means of face-to-face communication) has changed into the current situation (personalization by means of interactive media), and what consequences this has for the design of systems that aim at an audience of one.

### 1.3.1 The audience

Audience is the term that originally was used for the spectators in ancient Greek and Roman theaters and arenas, gathered to view a play or spectacle. Different kinds of events would attract different kinds of audiences, varying in, for example, education or social status. In the last 500 years, technological innovations have transformed the way in which we approach and perceive audiences, who have evolved from relatively small and homogeneous

groups of people into large and heterogeneous masses catered to by the mass media. This process primarily started in 1456 with the invention of printing, which allowed communicators to communicate their message to a larger and often unknown audience. Several centuries later, the industrial revolution and urbanization created a situation in which large geographically concentrated audiences could be reached more easily by means of newspapers and movie theaters. In the 1920s, the introduction of commercial broadcasting further reduced the limitations of the mass media's dependence on location. National radio shows, and a few decades later television shows, created nationwide audiences. Finally, the growing availability of Internet connections in the 1990s created the possibility for communicators to reach people, unconstrained by any geographical boundaries.

Creating one definition of "audience" to fit all the different strands of research that focus on addressing audiences is impossible (Webster, 1998). With this in mind, McQuail (1997) constructed a typology of "audiences" that spans the different research focuses. His typology classifies the research focuses on audiences by using a societal or a media perspective and subsequently a macro- or micro-level view.

On a macro-societal level, an audience is a group of people who can be considered a collective before their identification as an audience. An example of such an audience are the employees of an organization who are addressed through a company newsletter. The audience on a micro-societal level is the individual who chooses for himself or herself which TV program to "consume" or which Web site to visit. This view of the audience is central in the uses and gratifications theory, originally developed by Katz, Blumler, and Gurevitch (1973). According to the uses and gratifications theory, each media consumer consciously chooses the medium and message he or she wants to consume in order to fulfill a certain need (e.g., being informed of the latest news or being entertained).

McQuail's other perspective on audience, the media perspective, approaches people as a mass. On a macro level, a media audience consists of all the people who consume media content transmitted by one particular medium (e.g., the television audience or the book-reading public). More specific is the media audience on a micro level. This is the audience of one particular medium transmission. What binds these people is their consumption of a certain medium transmission (e.g., Monday night's eight o'clock news) and not their shared psychological or demographical characteristics.

The societal perspective on audiences can be characterized as a bottom-up perspective and focuses on the individual's motivations to consume certain media content or the small group's commonalities that makes them in-

teresting as a media audience. The media perspective is a top-down one. Instead of perceiving the individual or small group as the main party in the act of media consumption, the media perspective perceives the medium or a single transmission as the instigator of media consumption to which an audience is drawn. This perspective is prominent in media research and the design of media content (McQuail, 1997). In order to grasp commonalities among audience members, and to gear their communication towards these commonalities, players in the media analyze their audiences.

### 1.3.2 Analyzing the audience

The goal of audience analysis is "to identify its needs, document the perceived costs and benefits of addressing the needs, and formulate a program that addresses the needs in the most cost-beneficial manner to both the [receiver] and the [sender of the message]" (Lefebvre & Flora, 1988, p. 303). Napoli (2008) has outlined the evolution of audience analysis, a process strongly influenced by technological innovations. In the pioneering days of the mass media, audience analysis was performed by means of what Napoli calls the intuitive model: communicators applied their common sense and "gut feeling" to characterize their audience and to determine how it could be served best. After the Great Depression in the United States, the need for a better understanding of the audience arose as movies were becoming more expensive to produce and competition among media was growing. Therefore, a more systematic approach to audience analysis was applied. Sources such as box office figures, radio sales, or letters of complaint were used to deduce who was receiving the message and how it was appreciated. In the 1970s, the introduction of electronic information systems facilitated new ways of analyzing audiences. Large quantities of data could be easily collected (by means of sales systems or television set-top boxes), analyzed, and interpreted; and, as a result, a shift in focus took place. Instead of focusing on the number of people who had received a message and on their reception of the message, audience analysis increasingly focused on the demographics of the audience.

With the growing use of the Internet and the development of technologies like data mining, audience analysis has reached a whole new stage. The technological developments have provided an opportunity to collect data about individual audience members and to scrutinize their behavior at an extremely detailed level. It is, for example, possible to track and record an online bookstore customer's behavior via mouse clicks, viewing times, purchases, book ratings, etc. Subsequently, these data can be used to create a user model that states this user's tastes in literature, inferred on the basis of,

for example, owned books. In short, user modeling has made it possible to analyze audiences at a more detailed level than was possible before.

### 1.3.3 Targeting audience segments

As audience analysis was becoming a systematic undertaking, communicators—marketers in particular—realized that they could communicate more successfully if they addressed a small homogeneous segment of an audience instead of a large and heterogeneous population (Haley, 1968). In order to create advertisements that would have a higher persuasive effect with a specific subsection of the audience, Smith (1956) introduced "audience segmentation." Audience segmentation has been defined as "the process of identifying groups of customers who are relatively homogenous in their response to marketing stimuli, so that the market offering can be tailored more closely to meet their needs" (Brennan, Baines, & Garneau, 2003, p. 107). Audience segmentation, and the subsequent targeting of communication and product design at each segment, is done to find new, previously unaddressed target groups and to improve the communication to (potential) clients (Beane & Ennis, 1987). Ultimately, it has the potential to cater to the specific needs of customers and thus increase customer satisfaction and customer loyalty (Van der Geest, Jansen, Mogulkoç, De Vries, & De Vries, 2008). According to Kotler and Armstrong (1999), there are four kinds of data that can be used for audience segmentation:

1. Geographic data—e.g., similar country or city of residence
2. Demographic data—e.g., similar age, income or family size
3. Behavioral data—e.g., similar use of media or knowledge
4. Psychographic data—e.g., similar lifestyle or personality characteristics.

Although segmentation has been reported to be beneficial when marketing products, it has also been heavily criticized by scholars. The major criticisms of dividing an audience into segments are that there is no a priori segmentation approach that yields the best results, audience segments are often not discriminating and overlap, and, finally, segments are not stable, as people's characteristics and interests change constantly (Hoek, Gendall, & Esslemont, 1996). These drawbacks have led communicators to consider other ways of targeting their communication, mostly by focusing on individuals and addressing their unique characteristics, preferences, and contexts (Kara & Kaynak, 1997).

In the area of mediated communication, the possibilities of targeting communication at individuals have grown rapidly with the introduction of user modeling. Based upon a user model, a system can tailor output to each individual's unique needs, wishes, and context: personalization. Together

with user modeling, personalization changes the way in which communicators perceive and communicate with their audience. As a result, one can wonder what the importance and meaning of a concept like "audience" entails in this context. When the audience at large is replaced by a collection of individuals who are to be addressed with an individual message, do we even need a concept of "audience"?

### 1.3.4 Witnessing the end of the audience as we know it

Driven by advances in technology, the role of the individual audience member has transformed from a receiving party to the individual that is actively involved in the creation of a message. This shift is made possible by technological advances like hypermedia, cross-media, and user-generated content. Hypermedia has introduced a way of media consumption in which the individual audience member has gained control over the order in which content is consumed (Cover, 2006). And due to another innovation, cross-media, a message is not distributed by means of only one medium, but by different media that augment each other. For example, a television channel broadcasts a documentary about genetically modified rice after which a Web site facilitates a discussion on the topic between experts and viewers of the television broadcast. At the moment of writing, the latest development that has transformed the role of the audience is user-generated content (UGC). The Organisation for Economic Cooperation and Development (2007) has defined UGC as publicly available user content in which creative effort has been invested and that is created outside of professional routines and practices. Well-known examples of UGC collections are Flickr (www.flickr.com), where Web site visitors can place and tag (label) photos, and Wikipedia (www.wikipedia.org), a Web site where users can coauthor and coedit an encyclopedia.

Newly available technologies have enabled individuals to publish and personalize their own media content. As a result, the audience has transformed from a collective mass, traditionally addressed with one-way communication media, to unique individuals who are offered a more and more active role in the construction of a message (Livingstone, 2003; Tauder, 2005). This transformation is reflected in three changes in the traditional roles of communication senders and receivers and their relationships with each other (Bruns, 2007):
1. Senders do not consist of selected individuals or groups anymore, but of (a community of) different people with their own geographical location, knowledge, etc.

2. One person may assume different roles: generating the message at one moment, and consuming it at the other.
3. A message is continuously being created and is never finished.

These changes cast a new light on the traditional roles that senders and receivers have been allocated in communication theory in the past. People can be senders and receivers at the same time and later become receivers again. The roles of senders and receivers were conceived to be predefined and static, but are now dynamically assigned, depending on the task at hand. Communication has become a collaborative effort. As a result, professional communicators—and especially professional communicators working in the field of new media—should ask themselves whether they should still consider their target groups as audiences, as collective masses to be reached with one general message. Might it not be better to take a micro-societal view of the audience, the individual, and to reconsider the role of the individual in message construction and consumption?

The aforementioned changes in mediated communication make the term user more appropriate than audience member for characterizing the individual interaction with novel communication techniques like UGC and personalization. A user is an individual who can take on different communicative roles within one specific situation of use, like receiving and contributing content. In contrast, audience members are part of a mass, are primarily on the receiving end of communication, and are relatively passive during information consumption.

The shift of focus from a collective audience to individual users, served by personalization, requires a change in message design. The tools on which communicators have relied for decades are to be replaced; user modeling takes the place of audience analysis; and segmentation is put to its extreme in the process of personalization. As personalized messages are extremely sensitive to a correct correspondence with the individuals needs, wishes, and context (Kara & Kaynak, 1997), a heavy focus on the individual user throughout the design process is conditional (Canny, 2006). One way to ensure this correspondence is User-Centered Design.

## 1.4 User-centered design and personalization

In the mid-1980s, two publications introduced the User-Centered Design (UCD) approach (Gould & Lewis, 1985; Norman, 1986). In essence, UCD is a design approach in which the (prospective) user is the focus of attention and is consulted in all phases of the system design. In their landmark article, Gould and Lewis (1985) list three principles of UCD:

1. An early focus on users and tasks. Users should be consulted as early as possible, before system design, about their characteristics, needs, and wishes.
2. Empirical measurement. Studies should focus on actual user behavior and be conducted empirically.
3. Iterative design. Every substantial new version of the system should be tested with users, and the results of these studies should be incorporated in the next version of the system.

Later, they added a fourth principle, stating that systems should not be designed in isolation, but that all system aspects affecting usability (e.g., help functions or using multiple channels) should be designed in accordance and under one management body (Gould, Boies, & Lewis, 1991). These principles remain very abstract. In order to increase the practical value of the approach, Maguire (2001) divided the system development process into five phases:

1. Planning. In this phase the activities in the UCD process for a system are planned and geared upon each other
2. Context of use. In this phase the context of use of the prospective user is investigated
3. Requirements engineering. In this phase demands on the system design are elicited from relevant sources (e.g., prospective users) and translated into requirements.
4. Design. In this phase the system is designed.
5. Evaluation. In this phase the system is evaluated in order to get redesign input (formative evaluation) or to assess its effectiveness and usability (summative evaluation).

This development process should not be seen as a waterfall process in which phases are finished and not to be returned to. Instead, as stated in the third principle of UCD, the process is iterative and if necessary, designers should return to previous phases if the situation asks for it. For example, when a design team discovers that a requirement needs to be adjusted because of results of the formative evaluation, they should go back to the requirements phase.

So how is the UCD approach different for personalized systems? Traditionally, design has centered on abstractions of users, like audience segments or personas. System output had to comply with the needs, preferences, and contexts of these groups. When dealing with personalization, the design team's focus should be on the individual user. They have to ensure that personalized output is useful for every individual working with the per-

sonalized system in his or her unique context. Furthermore, the design team should focus on specific usability problems throughout the UCD process.

## 1.4.1 Identifying and preventing usability problems

Several authors have discussed how one can evaluate personalized systems. Gena (2005) and Gena and Weibelzahl (2007) have listed the methods that one can possibly apply during the UCD process of a personalized system. And although these overviews are a good reference point for the decision of which method to use at a given moment, they do not present a coherent approach in which multiple methods are used and geared toward each other. These overviews and several other publications, for example Höök (1997) and Weibelzahl (2005), have listed some pitfalls and ways to overcome them. The majority of these issues concern the design of a valid effectiveness measurement of a personalized system. The issue of applying UCD methods for understanding how users experience personalized output, and how this experience can be improved upon is rarely addressed in the literature.

A series of publications that give shape to the user experience with a personalized system has been written by Jameson (2003; 2007; 2009). Here, he lists seven usability issues that have a critical influence on users' satisfaction with personalization. These usability issues are not new, but with the rise of personalization, they have acquired a new meaning and increased importance. They are:

1. Predictability. Users must be able to predict the consequences of their actions for the generation of personalized output.
2. Comprehensibility. Users must be able to understand how user modeling and the tailoring of system output works.
3. Controllability. Users must be able to control their user model and the generation of personalized output.
4. Unobtrusiveness. Users must be able to complete their tasks without being distracted by personalization features.
5. Privacy. Users must not have the feeling that the generation of a user model infringes on their privacy.
6. Breadth of experience. Users must not lose the possibility of discovering something new because output only complies with their user model.
7. System competence. Users must not have the feeling that the system creates an invalid user model or does not personalize output successfully.

In order to ensure that a personalized system is designed such that it counters the possible negative effects of these issues, they have to be taken into account throughout the design process.

## 1.5 Thesis outline

The goal of this thesis is to contribute to the UCD toolkit for designers of personalized systems. Therefore, I will present four studies that provide either methodological implications or design guidelines, and span the different phases of the UCD process.

### Chapter 2: The role of trust and controllability in user acceptance of online content personalization

Chapter 2 focuses on the context of use phase in the UCD process. According to Maguire (2001), this is the moment to investigate the environment (technical, physical, as well as organizational) in which the technology will be used, the tasks that it must support and the users that will be using the new technology. Part of getting to know the users deals with understanding their attitudes towards the new technology. Do prospective users trust the new technology? Do they think it is an improvement over readily available technologies? User attitudes like these need to be understood by the design team and taken into account during the design of new technology. As a result, a new technology has a higher chance of user acceptance.

In this chapter, I report a large-scale web survey that has the goal to understand user acceptance of online content personalization, a popular form of tailoring website content. More specifically, the study focuses on the role of trust in the organization, trust in the technology and perceived controllability in the formation of the decision to (not) use this technology. These factors have been identified as important barriers to use personalization by several authors (Jameson, 2007; Pieterson, Ebbers, & Van Dijk, 2007).

### Chapter 3: User requirements engineering for a personalized social support e-service

Chapter 3 takes on the next phase in the UCD process: requirements engineering. User requirements engineering has been defined as "all the activities devoted to identification of user requirements, analysis of the requirements to drive additional requirements, documentation of the requirements as a specification, and validation of the documented requirements against the actual user needs" (Saiedian & Dale, 2000, p. 420). I show how user requirements for a personalized e-Service can be elicited and engineered, utilizing interviews with potential users, low-fidelity prototyping and

evaluation of this prototype. Furthermore, I will also demonstrate the added value of conducting these activities.

### Chapter 4: User-centered evaluation of personalized systems: A literature review

Chapter 4 is centered on the fifth phase in the UCD process: evaluation. It reports a literature review that gives an overview of published user-centered evaluations of personalized systems. It describes how these evaluations have been conducted and which lessons we can learn from them. Furthermore, it provides the reader with practical information on how to improve upon typical evaluation practice.

### Chapter 5: Identifying usability issues for personalization during formative evaluations: A comparison of three methods

Chapter 5 deals with the final phase in the UCD process, evaluation, as well. It reports on a study that compared the usefulness of three methods, concurrent thinking-aloud, interviews and questionnaires, for assessing usability issues for personalization (predictability, comprehensibility, etc.), as well as the perceived usefulness of personalization. This is done by evaluating a personalized internet meta-search engine with all three methods.

### Chapter 6: Reflection

In the final chapter of this thesis, I will first summarize the findings of the four studies. Then, I will reflect on some dominating views on personalization in the scientific literature and discuss how I think future design, evaluation and research should deal with these convictions. How does UCD align with a technical view on designing and evaluating personalization? Is personalization always better than technology that does not tailor to the individual? And what is the role of the user experience in design in relation with effectiveness and efficiency?

*In chapter 1, I have introduced the key concepts of this thesis: personalization and user-centered design. The first empirical chapter of this thesis deals with the first phase of the user-centered design process in which the (prospective) user is consulted: the context of use phase. Here, the design team needs to get to know the (prospective) users and their attitudes towards the new technology.*

*In the next chapter, I discuss a large-scale online experiment that has the goal to explore a set of these attitudes. More specifically, the study aims to investigate the role of trust and controllability in the formation of the decision to (not) use online content personalization, a popular form of tailoring content to an individual's characteristics, preferences and context. This knowledge can then be translated into user requirements for online content personalization in general.*

# Chapter 2

## The Role of Trust and Controllability in User Acceptance of Online Content Personalization

*"Never trust anything that can think for itself if you can't see where it keeps its brain."*
-- Arthur Weasley in Harry Potter and the chamber of secrets

## 2.1 Introduction

Content personalization is a form of personalization that is becoming a common practice on the World Wide Web. It takes many different forms: inserting information, removing information, altering fragments of text, re-arranging information, etc., all based on knowledge of the user (Knutov, De Bra, & Pechenizkiy, 2009). Based on the definition by the World Wide Web Consortium (Lewis, 2005), we define content personalization as the process of selecting, generating or modifying content units (e.g., text, pictures or video) in a given delivery context, based on user characteristics. If a visitor to a sports website, for example, only reads articles on soccer, the website may display new articles on soccer more prominently on its main page in the future. The goal of this technique is for people to more readily see, or be directed to, personally relevant information. This is especially relevant in large information databases (like a news website, an electronic learning environment or a digital museum catalogue). As a result, users can have a more efficient and satisfying experience with an information system. Tam and Ho (2006) have found that people find personalized content useful and are eager to explore personalized content further. Colineau and Paris (2009) found that people find the information they need more quickly when they can make use of content personalization. Content personalization techniques have been successfully applied in the context of personalized museum guides (e.g., Stock et al., 2007), online medical information (e.g., Cawsey, Grasso, & Paris, 2007) and content on mobile devices (D. Zhang, 2007).

Many factors influence a person's decision to use a specific technology like online content personalization. To provide a technology that is accepted by potential users, it is vital to determine which factors play a role in the formation of users' decisions and their relative importance. This knowledge can then be translated into design requirements. However, in the case of online content personalization and personalization in general, knowledge about the factors that shape users' decisions to accept this technology is scarce. Therefore, this study aims to answer the following research question:

What is the role of *trust in the organization, trust in the technology* and *perceived controllability* in the intention to use online content personalization?

In an online survey environment, 1,141 adult participants were shown scenarios describing a non-personalized e-government webpage with neighborhood information and one of four scenarios describing an adaptive or adaptable variant of the same page. The scenarios for personalized variants demonstrated different user modeling strategies. Finally, the participants completed an online survey.

## 2.2 Theoretical background

### 2.2.1 Content personalization: User-modeling strategies

The basis for personalization is always a user model: a file containing information about an individual's characteristics, preferences and context. This information is based upon interpretations of user or usage data. The system uses this interpreted data to infer what content is most suitable for an individual. Often, the collection of data about a user and the interpretation of this data are unobtrusive. When the user is not explicitly involved in the construction of the user model and the personalization of content based upon this model, a system is called *adaptive*. An example is Google adds. Other systems allow users to explicitly indicate how they want their personalized content to look; the personal BBC homepage (http://www.bbc.co.uk) is a well-known example. In this case, a system is *adaptable*.

The literature on audience segmentation mentions four types of data that can be used by a system to reason about users (Kotler & Armstrong, 1999). They are:
1. Geographic data: e.g., a person's country, city, or neighborhood of residence;
2. Demographic data: e.g., age, income, or family size;
3. Behavioral data: e.g., visited web pages or time spent on a web page;
4. Psychographic data: e.g., a person's social class, lifestyle, or personality.
After interpreting this data, inferences can be made. For example, a visitor to an online e-government service submits an e-form including the birth date February 2, 1937. The system interprets this entry as the user characteristic "65+" in the user model. Next, the inference is made that this person will be interested specifically in government information for seniors; hence, the user is shown this information on his/her personal myGovernment website.

The use of each subsequent type of data requires a more complex interpretation with a higher degree of uncertainty of a correct inference. An example of a simple interpretation and relatively straightforward inference was given above, in the example of government information for senior citizens. We will illustrate the complexity involved and the high degree of uncer-

tainty with an example utilizing psychographic data. Suppose that, on the basis of a completed personality questionnaire administered while logging in to a municipal website for the first time, the characteristic "leadership qualities" is inferred and stored in an individual's user model. Using this information, the website displays a recruitment text for a participatory council in the city. The interpretation in this example is very complex because several rules must be designed and applied to derive personal interests from questionnaire results. Consequently, the inference has a degree of uncertainty. After all, not all people who have leadership qualities will be willing to take a seat on a participatory council. An incorrect inference will likely result in user dissatisfaction as the content personalization is perceived as irrelevant or even erroneous.

The use of a different type of data may also affect users' concerns regarding their privacy. People are less likely to perceive the use of data as infringing their privacy when it is collected by a well-known organization, can be controlled by the individual, is perceived to be relevant for service provision, and can easily be used to make correct inferences about the individual (Culnan, 1993). According to several studies (Andrade, Kaltcheva, & Weitz, 2002; Malhotra, Kim, & Agarwal, 2004; Phelps, Nowak, & Ferrell, 2000), people are more hesitant to provide information that clearly describes what kind of person they are (such as hobbies) rather than simple factual information (such as age). This means that using each subsequent data type (geographic, demographic, behavioral and psychographic) as the basis for personalization is likely to be considered more privacy infringing, and *trust* will play a more important role.

Once user or usage data is interpreted and inferences are made, content can be personalized. As described in the introduction, this personalization can take many forms. Figure 2.1 displays the personalization process.



| User data | User model | Adapted content |
|---|---|---|
| Geographic data | Characteristics | Selection |
| Demographic data | Preferences | Layout |
| Behavorial data | Context | Modality |
| Psychographic data | Etc. | Etc. |

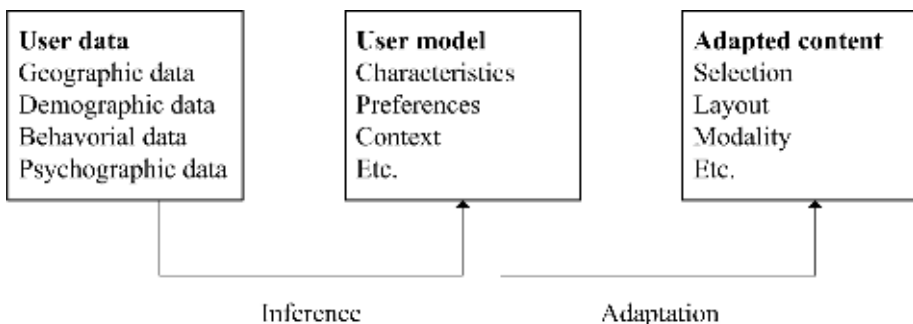Inference          Adaptation

Figure 2.1. The personalization process

### 2.2.2 User acceptance of personalized systems

As the rise of personalized systems has only been recent, the number of studies of the user acceptance of personalized systems is limited. Moreover, the kinds of systems that have been studied differ widely. They are as diverse as adaptive museum guides (Cramer et al., 2008; Pianesi, Graziola, Zancanaro, & Goren-Bar, 2009), an intelligent refrigerator (Rothensee, 2008), and a medical portal site (Pahnila, 2006). These studies uncovered a range of factors that contribute to a person's decision to use or not use a specific form of personalization (e.g., fun, perceived system control and perceived quality of system feedback). *Perceived usefulness* was found to be the most important factor in the case of the three systems mentioned. Jameson (2007) lists several usability issues that can have a negative effect on a user's experience of personalization, including *diminished predictability, infringement of privacy* and *diminished control*. This last issue, controllability, has also been identified by Kay (2006) as an important system characteristic that can hinder satisfaction with personalization. Another factor that is often named as a barrier to acceptance of personalization is a *lack of trust* (e.g., Chellappa & Sin, 2005; Pieterson, Ebbers, & Van Dijk, 2007). However, the identified factors seem to be eclectic and system- and situation-specific.

A related strand of research deals with consumers' reactions to the online collection of personal data for consumer profiling or audience segmentation. Factors that influence consumers' willingness to provide personal data for these goals include control over how their data is used (Graeff & Harmon, 2002; Olivero & Lunt, 2004), whether or not organizations share consumers' personal data with other organizations (Ackerman, Cranor, & Reagle, 1999), trust in Internet technology (Lusoli & Miltgen, 2009) and trust in an organization (Schoenbachler & Gordon, 2002).

The most significant frameworks for studying user acceptance are the Technology Acceptance Model (F. D. Davis, 1986), the Unified Theory of Acceptance and Use of Technology (Venkatesh, Morris, Davis, & Davis, 2003) and the Task-Technology Fit Model (Goodhue & Thompson, 1995). These models identify a limited number of factors to explain technology acceptance. For instance, the Technology Acceptance Model posits that the decision to use is affected by two factors: perceived usefulness and perceived ease of use. However, after assessing these factors and their influence on the decision to use a given technology, the model can only predict whether potential users will accept this technology or not. The motives and attitudes that lead to acceptance remain unknown (Baaren, Van de

Wijngaert, & Huizer, 2008). In other words, the model has low explanatory power. Determining the influence of system-specific factors on the decision to use seems a more useful approach to guide the design of a system. The advice that a system should be controllable, for example, is more helpful for designers than the advice that a system should be useful.

### 2.2.3 Trust and controllability

Based on previous studies, trust and controllability appear to be two important factors that determine whether a person will use personalization. Therefore, the present study will explore their role in the context of user acceptance of online content personalization.

*Trust* has been defined and operationalized very differently in the comprehensive literature on this topic. It can be approached one-dimensionally, as was done by Corritore, Kracher and Wiedenbeck (2003), who defined trust in the context of transactional or informational websites as "an attitude of confident expectation in an online situation of risk that one's vulnerabilities will not be exploited" (Corritore, Kracher, & Wiedenbeck, 2003, p. 740). Others researchers posit the concept of trust to be multi-dimensional. Trust is not one attitude but is the combination of different attitudes towards different concepts. By applying a fine-grained notion of trust, our grasp of users' motivation to trust a certain online service is better. Following Grabner-Krauter (2002), we will divide the concept of trust into *trust in the organization* and *trust in the technology*. Both forms of trust reflect an individual's willingness to be vulnerable towards someone or something.

*Trust in an organization* can be defined as "an individual's belief that an organization will fulfill a task for the individual with the individual's best interests in mind" (Mayer, Davis, & Schoorman, 1995, p. 712). In the context of online content personalization, this means that an individual allows an organization to determine what is useful information for him or her because he or she believes that this organization will not exploit this opportunity for causes that are not beneficial for the individual. *Trust in the technology* is defined as "an individual's belief that using a specific technology is safe and secure" (McKnight, Choudhury, & Kacmar, 2002, pp. 304-305). Technological structures (e.g., encryption) should instill confidence in the individual that using the technology will not cause harm, such as theft of personal data.

*Controllability* refers to a person's choice to be part of communication between two parties and the possibility of influencing the communication (Liu, 2003). It is an important aspect of interactive communication (Liu & Shrum, 2002). In the context of personalization, Jameson (2007) defined

controllability as "the extent to which the user can bring about or prevent particular actions or states of the system if she has the goal of doing so" (Jameson, 2007, p. 447). In other words, a user of a personalized system should have the option of influencing the coming about of personalized output. When a system provides personalized features, controllability becomes a crucial part of system usability (Jameson & Schwarzkopf, 2002). Especially in the case of adaptivity (where the user is not explicitly involved in the personalization process), it may be difficult, or even impossible, for users to influence this process. In a small-scale, qualitative study, Barkhuus and Dey (2003) found that perceived controllability decreases when systems make inferences about users without their involvement. Barkhuus and Dey also found that the perceived usefulness of the personalized features increases when users are infrequently consulted or not consulted during the personalization process. According to Godek and Yates (2005), this trade-off is only applicable in contexts where personalization helps users to select suitable information in a situation of information overload.

Table 2.1 displays the definitions of trust in the organization, trust in the technology and perceived controllability used in this study.

Table 2.1. Definitions of variables

| Factor | Definition | Based on |
|---|---|---|
| Trust in the organization (TO) | The belief that an organization will perform a particular action for an individual with the individual's best interests in mind. | Mayer, Davis & Schoorman (1995) |
| Trust in the technology (TT) | The belief that a technology has protective legal or technological structures (e.g., encryption) that assure that business can be conducted in a safe and secure manner. | McKnight, Choudhury & Kacmar (2002) |
| Perceived controllability (PC) | The belief that the user can choose to bring about or prevent particular actions or states of the system. | Jameson (2007) |
| Intention to use (IU) | The belief that a person will use a technology once it is available to him or her. | |

## 2.3. Experimental conditions and hypotheses

This study focuses on the role of trust and controllability in the acceptance of personalization. In line with the argument in Section 2.2.3, we distinguish between trust in the organization that provides the technology (TO) and trust in the technology (TT). In studies of system acceptance, acceptance of a technology has often been operationalized as the intention to use (IU). This is a variable that can be assessed before new technology is actually in use; it has been found to be a good predictor of the actual use of technology once it is available to users (Moon & Kim, 2001; Venkatesh, Morris, Davis, & Davis, 2003). Table 2.1 displays our definition of the intention to use.

It should be mentioned that, in this study, personalization has been defined as a *process*, not as the product, system output or web site that is the *outcome* of that process. This means that the manipulation in our experiment will not address website characteristics as such; rather, it is focused on presenting *approaches* to personalization that result in a particular system or website. The focus of our study is the formation of the acceptance of these approaches.

### 2.3.1 Experimental conditions

We assessed the role of trust in the organization (TO), trust in the technology (TT), and perceived controllability (PC) for the formation of users' intention to use (IU) by means of an online survey that presented four possible approaches to online content personalization by a fictive municipality as well as a non-personalized baseline condition. These five experimental conditions are as follows:

§ Condition 1 (control): **No personalization**: every user sees the same, non-personalized homepage.
§ Condition 2: **Adaptable** approach: users can determine which pieces of information are displayed on their personal homepage.
§ Condition 3: **Adaptive/demographic** approach**:** user characteristics derived from demographic data determine the selection of information displayed on users' personal homepage.
§ Condition 4: **Adaptive/behavior** approach**:** user characteristics derived from behavioral data determine the selection of information displayed on users' personal homepage.
§ Condition 5: **Adaptive/psychographic** approach: user characteristics derived from psychographic data determine the selection of information displayed on users' personal homepage.

The three adaptivity approaches are based on the different kinds of data that can be used for user modeling (as listed in Section 2.2.1). We have not included adaptivity based on geographic data because it is very similar to demographic data.

### 2.3.2 Hypotheses

The variable *Trust in Organization (TO)* was operationalized as "Trust in the Municipality" because this study used the case of online content personalization provided by a municipal website. This context was chosen because a municipality is likely to have many sources of data on which to base content personalization and because municipalities have to provide information for a wide range of people, which allowed us to approach a wide selection

of participants. Because the municipality in our experiment was fictive, we assessed the participants' trust in the municipality in which they actually live. Because participants had probably interacted with this organization (by visiting the website, applying for a passport, etc.), we think this is a more valid measurement than asking participants to express their trust in a fictive municipality of which they have only seen several website screenshots. Trust in the government has been found to affect the intention to use (IU) e-Government initiatives positively (Bélanger & Carter, 2008). We hypothesize that this finding also holds for municipal (personalized) content provision.

H1: Trust in the organization (TO) is positively related to the intention to use both non-personalized and personalized approaches to online content provision.

Previous research has shown that trust in the safety of the internet in general (McKnight, Choudhury, & Kacmar, 2002) or e-Services specifically (Kim & Kim, 2005) positively affects the IU electronic services. We assume that this influence of *Trust in the Technology (TT)* on IU also holds for both the non-personalized and personalized approaches.

H2: Trust in the technology (TT) is positively related to the intention to use both non-personalized and personalized approaches to online content provision.

*Perceived controllability (PC)* has been found to be a factor that positively influences a person's intention to use e-Services (Lee & Allaway, 2002). On the basis of this finding, we expect PC to influence IU.

H3: Perceived controllability (PC) is positively related to the intention to use both non-personalized and personalized approaches to online content provision.

Finally, it is very likely that the *importance* of different factors for the formation of IU differs with the different approaches to online content personalization. For example, TO might be more important when information about a person's lifestyle is collected and used to personalize output (adaptive/psychographic) than when users can customize their own website (adaptable). No previous research has delved into this matter, which makes it difficult to formulate hypotheses for these differences. Therefore, we will

apply an explorative approach to determine the relative importance of the factors in the IU for the different approaches to content provision.

## 2.4. Method

To test our hypotheses, we conducted an online experiment in combination with an online survey. First, participants received a short introduction. Next, they were asked to rate their agreement on four items that assessed trust in the organization (TO). After they were shown the no-personalization scenario, the participants were randomly guided to one of the four personalization conditions or directly to the questionnaire. The questionnaire included statements on trust in the technology (TT), perceived controllability (PC) and the intention to use (IU), offered in a random sequential order. Finally, the participants were asked to answer questions about their demographical characteristics.

### *2.4.1 Scenarios*

Using scenarios supplemented with screenshots, we presented very simple prototypes of the different (personalized) approaches to content provision to our participants. Such prototypes can elicit user opinions on technology acceptance factors that resemble the opinions that are elicited when people interact with the technology (F. D. Davis & Venkatesh, 2004).

Each scenario consisted of a short narrative in which the participant was told about a fictitious person (Peter) who uses the website of his municipality (the fictive municipality of *Grootstad*, Dutch for *Bigcity*) to gather relevant news or information about his neighborhood (*Waterwijk)*. This scenario was supplemented by screenshots depicting the workings of the (personalized) approach to content provision on the Grootstad website. Because evaluation participants often find it difficult to notice tailored output when confronted with personalization (Weibelzahl, 2005), we showed the participants not only what Peter's personal website would look like in the personalization conditions, but also what Karin (another fictitious person) would see on her personal website. This way, we could be sure that the participants would notice and understand our experimental manipulations.

All scenarios described visiting a page on the website that listed neighborhood information. This information consisted of several snippets (e.g., building permits issued in the neighborhood or a calendar showing the collection of garbage). The topics of these snippets were consistent to rule out an effect of information of varying usefulness in different scenarios. The

exact content of each snippet was altered to align with the possibilities of each approach to personalization.

Each scenario describing a personalized form of content provision first displayed and explained the login procedure that the fictive website required. The login procedure utilized an authentication procedure called DigiD (http://www.digid.nl/english/), the standard authentication procedure for Dutch governmental websites. Next, participants were provided with one of the scenarios that described and showed a personalized approach to content provision. A summary of the different scenarios is as follows:

§ Condition 1: **No personalization**. Participants were first shown the homepage of the Grootstad website and then told about and shown the page of the Grootstad website that provided information about a neighborhood in this city. In this condition, the neighborhood page supplied one-size-fits-all information, like recently issued building permits in the whole neighborhood. This condition serves as a baseline comparison for the other conditions. Every participant was shown the no-personalization scenario to make the difference between standard and personalized content provision explicit.

§ Condition 2: **Adaptable**. Participants were told about and shown the same page with neighborhood information. Now, however, they were also informed of the option to explicitly choose the topics they would like to receive information about on this page, such as news about cultural activities in the neighborhood or a list of recently issued building permits. They were also informed of the option to change their decisions at a later time.

§ Condition 3: **Adaptive/demographic**. Participants were told about and shown the neighborhood page that was constructed based on the fictive person's demographics. For example, the homepage showed only announcements of issued building permits in a radius of 250 meters of Peter's address.

§ Condition 4: **Adaptive/behavior**. Participants were told about and shown the neighborhood page that was constructed based on the fictive person's previous behavior on the website. For example, participants were told that in the past, the fictive person, Peter, reported to the municipality on his online tax form that he owned a dog. As a result, the neighborhood page included news about places in the neighborhood where people are allowed to let their dogs run free.

§ Condition 5: **Adaptive/psychographic**. Participants were told about and shown the neighborhood page that was constructed based on the fictive person's psychographic data. First, participants were shown a screenshot

of a page on which Peter was asked to rate his agreement with several statements on Likert scales to ascertain which kind of predefined life-style suited him best (taken from the VALS framework and survey for audience segmentation based on psychological traits and key demographics (Strategic Business Insights, 2009)). Then, participants saw a screenshot that displayed the result of the fictive person's lifestyle test. Peter, for example, was typed as an innovator (a lifestyle type in the VALS framework): someone who has an active lifestyle and likes to take charge. Finally, participants were shown the personalized neighborhood page that included, among other snippets of information, the recruitment text for a position in a participatory council in the neighborhood.

Exemplary screenshots can be found in Appendix A.

### 2.4.2 Survey items

The survey items can be found in Appendix B. To ensure high construct validity, we adapted measurement scales that have proven their value in past studies. The items that measure TO are adapted from Bélanger and Carter (2008) and are specifically focused on trust in the municipality (which is the focus of our demonstration material), while the items assessing TT are based upon McKnight, Choudhury and Kacmar (2002). The items that determine PC are derived from Liu (2003). Our IU scale is based on Davis (1989) and expanded with one item from Gefen, Karahanna and Straub (2003) and one item of our own. We (re-)worded several items negatively. All statements were accompanied by a 7-point Likert scale, ranging from completely disagree (1) to completely agree (7).

### 2.4.3 Pretest

The scenarios and questionnaire were pretested before they were deployed. Six men and four women (ages ranging from 26 to 70) were shown all of the scenarios and completed the questionnaire for one approach. Hence, the questionnaire was completed twice for each approach. The pretest participants were asked to comment on anything they found unclear or when they found it difficult to answer a question. As a result of the pre-test, the text of the scenarios underwent minor changes. We added the option to answer "I don't know" to each TO item, and one item was rephrased.

The pretest also served as a manipulation check. After they read the scenarios and looked at the screenshots, we asked the participants to explain how the neighborhood information was generated. It turned out that the participants had no trouble understanding and retelling how each approach to

content provision worked. Hence, translation of the different content provision approaches into the scenarios and screenshots can be considered successful.

### 2.4.4 Recruitment of participants

Participants were recruited via two commercial online research panels. The only restriction we placed on their selection was that participants had to be 18 years of age or older. This way, we could be sure the participants had some experience with information about their neighborhood. Both panels selectively chose participants to generate a representative sample of the Dutch population. They supplied approximately the same number of participants. In Panel 1, participants were rewarded for their time with credit points that could be exchanged for gifts in an online store. In Panel 2, participants had a chance of winning a gift voucher for participating.

## 2.5 Results

### 2.5.1 Participant demographics

Participants who completed only a small part of the survey (e.g., abandoning the survey after the first page of questions) were removed. In total, 1,141 people completed our online survey. A response rate could not be calculated. Men accounted for 54.0% of the participants, and 46.0% of the participants were female. Furthermore, 31.4% of the participants were 18 to 40 years old, 55.5% were 40 to 65 years old and 13.1% were older than 65 years. The majority of the respondents completed education at the intermediate vocational level (47.9%) followed by completion at the higher vocational level (24.1%), lower vocational level (17.2%) and, finally, the university level (10.8%). The participants used the Internet on a daily basis (92.6%) or used it three to five times a week (6.3%). Finally, we assessed the frequency with which they visited their municipality's website. Most participants visited this website a few times a year (49.1%). Two groups of about equal size visited this website (almost) never (21.6%) or once or twice a month (22.5%). Only 6.7% visited the website of their municipality once a week or more. In all, despite a small overrepresentation of participants who completed education at a higher vocational level, the participants were representative of the Dutch population (Statistics Netherlands, 2010). The participants were well spread across the five conditions (no personalization, adaptable, adaptive demographics, behavioral and psychographics), with group sizes of 223, 221, 220, 246 and 231, respectively.

### 2.5.2 Assessing measurement quality

The first step in our analyses was to assess the measurement quality of our online survey and, if necessary, to improve it. After rescaling the negatively worded items, we assessed an initial Cronbach's alpha score for each factor in every condition and the item-total correlations. Results of these analyses can be found in Appendix C.

Three negatively worded items (TO4, TT1 and PC3; abbreviations refer to Appendix C) appeared to be troublesome. Participants may have had difficulty forming a response because negative wording increases the difficulty of the item, which could have led to deviant answering behavior (Fowler Jr., 1995). These items needed to be removed from the measurement scale in one or more conditions to generate reliable measurement scales. Because we wanted to compare the factors between the different conditions at a later stage in the analyses, we decided to remove these three items from every measurement scale in every condition.

When we compared the factor loadings of the items, we noticed that the third trust in the technology (TT) item ("Your personal data are protected well when using this page") had a far lower item-total correlation in the baseline condition than in the other conditions. In the scenario and prototype of the baseline condition, no personal data were required from the fictive user. As a result, participants may have had problems responding to this statement, and the item did not measure the same construct in the different conditions. Therefore, we removed this item from the measurement scale in every condition. As a result, TT was assessed by means of two items.

The resulting Cronbach's alphas of the measurement scales can be found in Table 2.2 and are all above the minimum level of .7 (Nunnally & Bernstein, 1994), while most Cronbach's alphas are good to excellent (higher than .8 or .9). Cronbach's alpha could not be determined for the TT factor because it consisted of only two items. Instead, we calculated correlations between the two items. In the five conditions, $r$ is respectively .51, .56, .53, .70 and .48 (all significant at $p < .01$). These numbers appear to indicate that the two items assess the same underlying concept.

Our next step was to check for multicollinearity. We scanned the correlation matrices of the five conditions for problematic correlations. All correlations were below .8, which suggests there were no problems with multicollinearity; each measurement scale assessed a separate psychological factor.

Table 2.2. Construct reliability: Cronbach's α

| | 1. No personal-ization | 2. Adaptable | 3. Adaptive/ demographic | 4. Adaptive/ behavior | 5. Adaptive/ psychographic |
|---|---|---|---|---|---|
| Trust in the organization | .86 | .86 | .89 | .91 | .93 |
| Trust in the technology | n.a. | n.a. | n.a. | n.a. | n.a. |
| Perceived controllability | .75 | .75 | .76 | .82 | .84 |
| Intention to use | .89 | .93 | .89 | .94 | .95 |

The analyses, described in this section, improved the reliability of the factor measurements. This allowed us to proceed with our next step, namely to describe and compare the participants' perceptions of trust in the organization, trust in the technology and controllability, as well as their intention to use, for the different approaches. The comparisons served as a manipulation check; they clarify whether the participants experienced the five approaches differently. For example, the adaptable condition explicitly provides users with the tools to control the personal selection of information, which should be reflected in a higher mean score for the perceived controllability scale.

### 2.5.3 Trust in the organization

Table 2.3 shows the mean scores and standard deviations for trust in the organization (TO) in each condition, while the boxplot in Figure 2.2 displays the distribution of the participants' responses. In all conditions, the participants indicated that they trusted their municipality or had a neutral disposition towards this issue. On average, 10.6% of the participants utilized the "I don't know" option for any of the TO items.

Table 2.3. Variables: means and standard deviations (S.D.), assessed on a seven-point scale

| | 1. No personal-ization | | 2. Adaptable | | 3. Adaptive/ demographic | | 4. Adaptive/ behavior | | 5. Adaptive/ psychographic | |
|---|---|---|---|---|---|---|---|---|---|---|
| | mean | SD | mean | SD | mean | SD | mean | SD | mean | SD |
| Trust in the organization | 4.86 | 1.12 | 4.83 | 1.12 | 4.80 | 1.18 | 4.70 | 1.21 | 4.86 | 1.17 |
| Trust in the technology | $4.71^{3/4/5}$ | 1.03 | $4.59^{3/4/5}$ | 1.14 | 4.26 | 1.11 | 4.23 | 1.42 | 4.10 | 1.21 |
| Perceived controllability | $4.52^{3/5}$ | 1.04 | $5.04^{1/3/4/5}$ | .93 | 4.19 | 1.13 | $4.35^{5}$ | 1.26 | 4.01 | 1.27 |
| Intention to use | $4.84^{5}$ | 1.18 | $5.83^{5}$ | 1.34 | $4.61^{5}$ | 1.31 | $4.53^{5}$ | 1.50 | 3.85 | 1.60 |

Note: numbers in superscript behind mean indicate that this mean is significantly higher than the mean of the same variable of the condition with the following number:
1 = No personalization; 2 = Adaptable; 3 = Adaptive/demographic; 4 = Adaptive/behavior; 5 = Adaptive/psychographic

Next, we assessed whether trust in the organization differed over the conditions by means of a univariate ANOVA analysis. The approach to online content provision did not affect TO ($F (4, 1092) = .72, p > .05$). In each condition, TO was appreciated the same. Because TO could not be influenced by our scenarios and screenshots (it was asked at the start of the experiment), this result does not come as a surprise; rather, it shows that the participants were well spread over the different conditions.
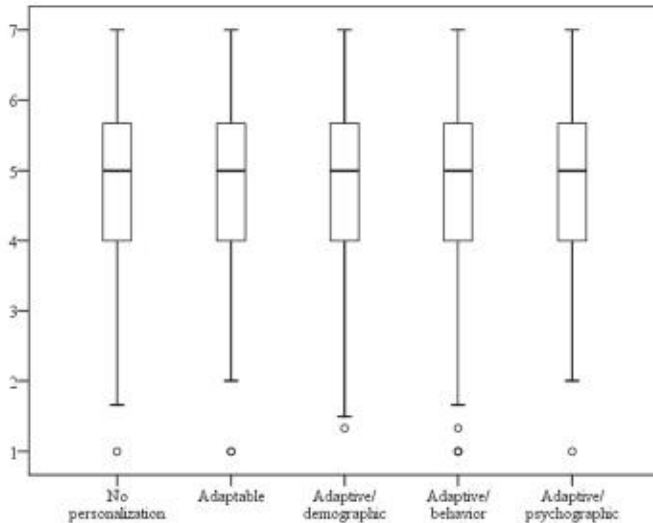


Figure 2.2. Boxplot for trust in organization

### 2.5.4 Trust in the technology

Table 2.3 and the boxplot in Figure 2.3 show that in the no-personalization and adaptability conditions, the majority thought the technology was safe.

In the conditions in which content was implicitly tailored to the users' context, however, perceptions were more dispersed. In the adaptivity/demographic condition, participants showed some trust in the technology or valued Trust in the Technology (TT) at or slightly below the neutral point. The scenarios and screenshots demonstrating the adaptive/behavior approach evoked a wider range of responses (as can be seen in Figure 2.3), ranging from a positive appreciation of TT to an appreciation at or slightly below the neutral point. Finally, adaptivity/psychographic resulted in a small range of answers, from slightly above to slightly below the neutral point.
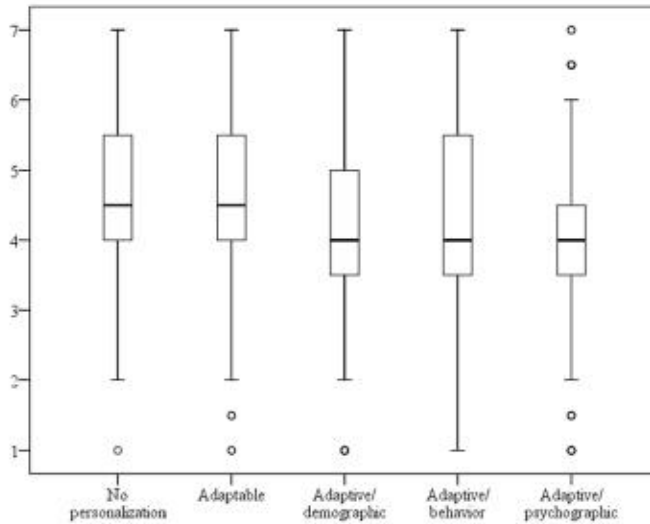
Figure 2.3. Boxplot for trust in technology

A univariate ANOVA analysis revealed that TT differed over the conditions: $F(4, 1092) = 10.91$, $p < .001$. Next, we conducted a post hoc Bonferroni test with a 5% significance level to determine exactly which conditions differed. Results can be found in Table 2.3. The no-personalization and adaptability condition both induced a higher TT than any of the conditions that involved an adaptive approach to online content personalization. This result suggests that participants understood that data was implicitly collected and interpreted in the adaptive conditions.

### 2.5.5 Perceived controllability

Table 2.3 and the boxplot in Figure 2.4 show that in three conditions (no-personalization and adaptivity based upon demographics or previous behavior), Perceived Controllability (PC) was somewhat positive to neutral.

In the adaptability condition (where the users were given the most tools to control the selection of information), PC was appraised as positive to somewhat positive. Finally, in the adaptivity/ psychographic condition, PC was valued from somewhat positive to somewhat negative.

A univariate ANOVA analysis indicated that the evaluation of PC differed over the five conditions: $F(4, 1092) = 25.39$, $p < .001$. By means of a post hoc Bonferroni test and a 5% significance level, we distinguished between the conditions. Results can be found in Table 2.3. Not surprisingly, PC was valued higher in the adaptability condition than in any of the other conditions, probably due to the inclusion of features providing explicit con-
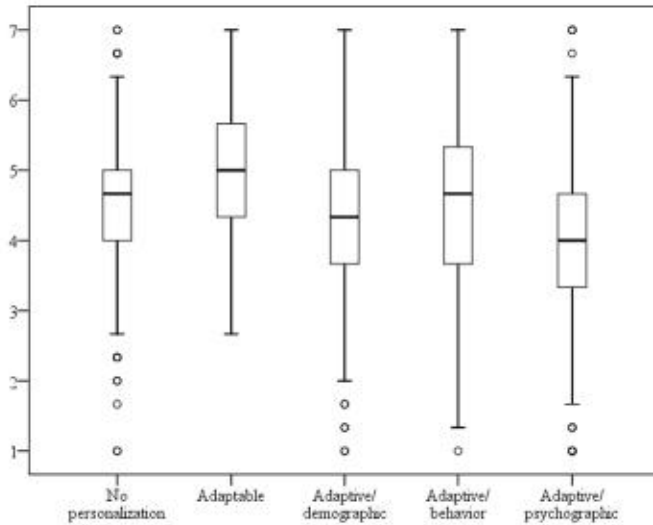
Figure 2.4. Boxplot for perceived controllability

trol of the content selection. The scenario and screenshots of the adaptivity based upon demographic and psychographic conditions were perceived to be less controllable than the prototype of the no-personalization condition. Finally, PC was valued lower in the condition of adaptivity based upon psychographics than in the condition of adaptivity based upon previous behavior. These results indicate that implicit collection and interpretation of personal data and the specific adaptive approach affected participants' perceptions of controllability.

### 2.5.6 Intention to use

As for the previous variables, means and standard deviations for the scores awarded to the intention to use (IU) can be found in Table 2.3. The boxplot depicting the distribution of responses can be found in Figure 2.5.

For all conditions except the adaptivity based upon psychographics condition, IU ranges from positive to neutral. In the case of the adaptivity based upon psychographics conditions, responses differ widely (as is reflected in the high standard deviation) and range from a somewhat positive intention to use to no intention to use.

We tested whether IU differed over the five approaches to content provision by means of a univariate ANOVA analysis. This appeared to be the case: $F(4, 1092) = 18.37$, $p < .001$. Results can be found in Table 2.3. The only condition for which IU differed was the adaptivity based upon psycho-

graphics condition. Here, IU was significantly lower than IU in all other conditions.
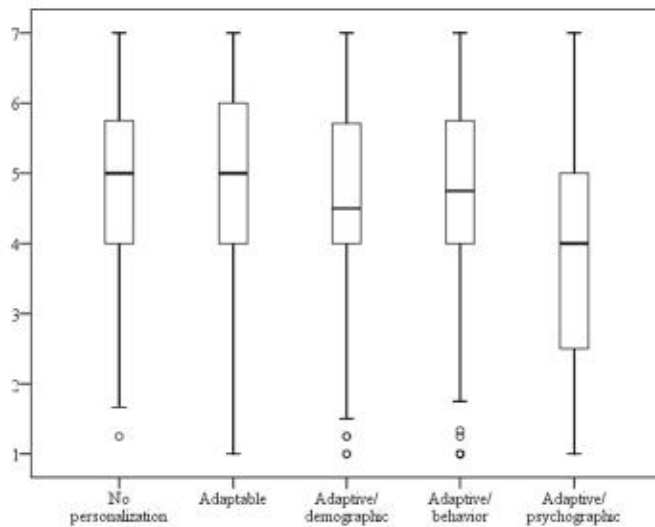


Figure 2.5. Boxplot for intention to use

## 2.5.7 The formation of the intention to use

We expected the Intention to Use (IU) online content provision to be partly determined by Trust in the Organization (TO), Trust in the Technology (TT) and Perceived Controllability (PC). Furthermore, we wanted to assess the relative importance of these factors over the different conditions for the formation of the intention to use. For example, is TT more important in adaptive approaches to online content personalization than in the adaptable approach?

We analyzed the influence of TO, TT and PC on IU for each condition using separate multiple regression analyses. Because we have not found any indication in previous studies of the relative importance of these factors, the independent variables TO, TT and PC were included in the analyses by means of backward stepwise regression. Tables 2.4 to 2.8 display the results of the regression analyses for the five conditions. In all the conditions, TO did not have a significant influence on IU and was removed from the model. TT and PC were retained because they did significantly influence IU.

PC turned out to have the greatest influence on IU in each condition, while TT had a moderate effect on IU. The exception here is the adaptive/behavior condition, in which the influence of TT on IU was slightly higher than the influence of PC on IU. The high $R^2$ values for IU (ranging

from .43 to .60) are remarkable given that IU is explained by only two factors (TT and PC). These numbers underline the great importance of trust in the technology and perceived controllability when designing systems that provide online content personalization.

Table 2.4. Regression analyses: No personalization condition; Dependent variable: IU

|  | B | SE B | β |
|---|---|---|---|
| Step 1 |  |  |  |
| Constant | .90 | .35 |  |
| Trust in the organization | .011 | .06 | .01 |
| Trust in the technology | .29 | .07 | .25* |
| Perceived controllability | .57 | .07 | .51* |
| Step 2 |  |  |  |
| Constant | .93 | .30 |  |
| Trust in the technology | .29 | .07 | .26* |
| Perceived controllability | .57 | .07 | .51* |

Note: R² = .46 for Step 1, Δ R² = .00 for Step 2 (n.s.).
* p < .001

Table 2.5. Regression analyses: Adaptable condition; Dependent variable: IU

|  | B | SE B | β |
|---|---|---|---|
| Step 1 |  |  |  |
| Constant | -.43 | .42 |  |
| Trust in the organization | -.03 | .06 | -.02 |
| Trust in the technology | .33 | .07 | .28* |
| Perceived controllability | .77 | .09 | .53* |
| Step 2 |  |  |  |
| Constant | -.50 | .38 |  |
| Trust in the technology | .33 | .07 | .27* |
| Perceived controllability | .76 | .08 | .53* |

Note: R² = .50 for Step 1, Δ R² = .00 for Step 2 (n.s).
* p < .001

Table 2.6. Regression analyses: Adaptive/demographic condition; Dependent variable: IU

|  | B | SE B | β |
|---|---|---|---|
| Step 1 |  |  |  |
| Constant | .78 | .36 |  |
| Trust in the organization | .04 | .06 | .04 |
| Trust in the technology | .33 | .07 | .28* |
| Perceived controllability | .53 | .07 | .46* |
| Step 2 |  |  |  |
| Constant | .91 | .31 |  |
| Trust in the technology | .34 | .07 | .29* |
| Perceived controllability | .53 | .07 | .46* |

Note: R² = .43 for Step 1, Δ R² = .00 for Step 2 (n.s.).
* p < .001

Table 2.7. Regression analyses: Adaptive/behavior condition; Dependent variable: IU

| | B | SE B | β |
|---|---|---|---|
| **Step 1** | | | |
| Constant | .68 | .31 | |
| Trust in the organization | -.04 | .06 | -.03 |
| Trust in the technology | .49 | .06 | .46* |
| Perceived controllability | .46 | .06 | .39* |
| **Step 2** | | | |
| Constant | .54 | .25 | |
| Trust in the technology | .47 | .06 | .45* |
| Perceived controllability | .46 | .06 | .39* |

Note: $R^2 = .55$ for Step 1, $\Delta R^2 = .00$ (n.s.).
* $p < .001$

Table 2.8. Regression analyses: Adaptive/psychographic condition; Dependent variable: IU

| | B | SE B | β |
|---|---|---|---|
| **Step 1** | | | |
| Constant | -.74 | .35 | |
| Trust in the organization | .03 | .06 | .03 |
| Trust in the technology | .49 | .07 | .37* |
| Perceived controllability | .61 | .07 | .49* |
| **Step 2** | | | |
| Constant | -.62 | .26 | |
| Trust in the technology | .49 | .07 | .37* |
| Perceived controllability | .61 | .07 | .49* |

Note: $R^2 = .60$ for Step 1, $\Delta R^2 = .00$ (n.s.).
* $p < .001$

The results of our regression analyses do not support H1: trust in the organization is positively related to the intention to use for both non-personalized and personalized approaches to online content provision. TO did not affect IU in any of the conditions.

We did find evidence for hypothesis H2: trust in the technology is positively related to the intention to use for both non-personalized and personalized approaches to online content provision. In every condition, TT was found to have a reasonable or large effect on IU (β ranging from .26 to .45).

Our last hypothesis, H3 (perceived controllability is positively related to the intention to use both non-personalized and personalized approaches to online content provision), was also supported by our results. More specifically, for every approach except for adaptive/behavior, PC was found to be the most important antecedent of IU (β ranging from .39 to .53).

## 2.6 Conclusions and discussion

The results of this study show that, overall, *perceived controllability* is an extremely important antecedent of the intention to use online content personalization. The approach to online content personalization that is perceived to be most controllable offers users the option to explicitly state the content they want to see (adaptability). *Trust in the organization* providing online content personalization appeared, at least in this study, not to play a role in the formation of the intention to use for any form of online content personalization. *Trust in the technology*, on the other hand, had a moderate effect on the intention to use any form of online content personalization. It is noteworthy that trust in the technology is more important for adaptive approaches to online content personalization than for an adaptable or non-personalized approach, while at the same time users' trust in the adaptive technology is relatively low.

Because we acquired a very high degree of explained variance for the intention to use the approaches to online content personalization, we can state that trust in the technology and, especially, perceived controllability are crucial when designing online content personalization. Users want to be able to control the tailoring of the information provided to the individual. This can be done by *applying an adaptable approach* to online content personalization in which the users can select the (types of) information they want to see on their personal pages. Another solution is to provide users the option to *view and alter the user model* generated by an adaptive-approach system to content personalization. Given our finding that the adaptable approach induces higher trust in the technology, we think that providing this approach to online content personalization is likely to have the highest chance of acceptance by users.

In addition to being controllable, the technology behind online content personalization should also be perceived as trustworthy. This means that users must feel that the technology with which they are interacting is safe and that storage of personal data is secure. Several overviews of guidelines for *designing for trust and security* can be used for this goal. Egger (2003), for example, has listed and validated many guidelines for inducing trust in online technology. They include "complement browser feedback with text to inform users that they are on a secure page" and "be audited by and display the seals of an independent trusted third party" (Egger, 2003, p. 54).

In this study, trust in the organization did not have a significant effect on the intention to use for any kind of online content personalization. The significant variance in the participants' reactions to the related items (see Fig-

ure 2.2) suggests that trusting or not trusting the organization does not lead to a higher or lower intention to use online content personalization. Consequently, our results imply that implementing design cues that can increase trust in the organization will probably not have an effect on the use of online content personalization. However, a remark must be made about the generalizability of this conclusion. In this study, trust in the organization was assessed for one specific kind of organization: the municipality in which the participant lived. The role of trust in the organization on the intention to use may be different for government organizations than for commercial organizations that offer online content personalization. These organizations have different motives (non-profit versus profit) and, as a result, people may estimate the risk involved in interacting with these types of organizations differently (Awad & Krishnan, 2006; Bélanger & Carter, 2008; Beldad, De Jong, & Steehouder, 2010). Future research must determine whether trust in the organization affects the intention to use online content personalization in a commercial setting. Furthermore, trust in the organization was assessed for the municipality in which the participant lived and not for the fictitious municipality that provided the online content personalization. As previously stated, we think this is a more valid measurement of trust in the organization. Participants have more experience with their municipality than with an organization of which they only have seen website screenshots. As such, we used trust in the participants' current municipality as a proxy variable for trust in the fictitious municipality offering the technology. Future research should confirm the lack of an influence of trust in the organization on the intention to use online content personalization for real-world settings.

Designers must keep in mind that opting for a specific kind of online content personalization carries specific design requirements. While controllability appeared to be a crucial aspect of design for any type of content personalization, trust in the technology became more of an issue when an adaptive approach was applied and the system unobtrusively collected and interpreted user data. It is possible that other usability issues that are important when designing content personalization (such as system predictability or the inability to discover new, unexpected things; for overviews, see Jameson (2007, 2009)) should be treated differently for different kinds of online content personalization. Future research should explore the role of these issues in relation to user acceptance of different kinds of online content personalization.

*In the previous chapter, we found that perceived controllability and trust in the technology play an important role in the formation of the decision to (not) use online content personalization. This knowledge can be translated into user requirements for this technology in general and as such, serve as input for the follow-up phase in the design process.*

*The elicitation, formulation and evaluation of system-specific requirements is the focus of the third phase of the user-centered design process: requirements engineering. Chapter 3 proposes a user-centered approach to requirements engineering for personalized e-Government services and validates the approach by means of a case study. The use of this approach allows user-centered designers to incorporate user input in the design of personalized e-Government services. Ultimately, this is likely to increase the fit between the final e-Service and user characteristics, preferences and context.*

# Chapter 3

## User Requirements Engineering for a Personalized Social Support e-Service

This chapter is an adaptation of
Van Velsen, L., Van der Geest, T., Ter Hedde, M. & Derks, W. (2009). Requirements engineering for e-Government services: A citizen-centric approach and case study. *Government information quarterly, 26*(3), 477-486.

## 3.1 Introduction

Providing citizens with one-stop personalized electronic government services is considered to be the kind of service provision that every government organization should strife for (Andersen & Henriksen, 2006). These kinds of services should make it easier for citizens to apply for, and manage the services they need. By applying a UCD approach, personalized e-Government services could live up to this goal. However, the actual e-Services that government agencies have provided in the last few years have fallen short of being user-centered (Soufi & Maguire, 2007) due to a lack of representative user involvement in the design process (Følstad, Jørgensen, & Krogstie, 2004). A survey among the most innovative European e-Government service designers showed that they mainly consult users when evaluating prototypes (Benchmark personalization of governmental eServices for citizens, 2008). Such a design tactic is out of step with the principles of UCD in which repeated consultancy of the prospective users from an early stage in the system design process onwards is advocated (Gould & Lewis, 1985). In order to design high quality personalized e-Government services that comply with the needs and wishes of citizens, a user-centered design approach needs to be developed within this context. The approach should not only include activities that deal with the evaluation of prototypes, but also, and perhaps even more importantly, activities deployed during the requirements engineering stage.

Several studies have shown the added value of including a user-centered requirements engineering stage in the system development process: by involving prospective users, requirements gain in accuracy (Damodaran, 1996; Kujala, 2003). But a user-centered requirements engineering approach also brings positive effects over time. According to Kujala (2003), it prevents the inclusion of superfluous features and increases system acceptance. Ultimately, user involvement leads to increased usability (Karat, 1994) and usefulness (Mao, Vredenburg, Smith, & Carey, 2005) of the final system. From a cost-benefit perspective, user involvement in the requirements engineering stage is also interesting: it can save money because potential problems can be fixed early on (Karat, 1994).

Currently, the literature shows a lack of publications that deal with the intricacies of user requirements engineering for personalization. Gena and

Weibelzahl (2007) have published a concise list with methods that can be applied in this stage. However, they do not discuss the methods in relation to each other and as a result, the reader is not provided with a coherent approach to user requirements engineering in this context. In Van Velsen, Huijs and Van der Geest (2008) it is discussed how requirements for a personalized enterprise resource planning system can be elicited from future users successfully, but not how these requirements should be validated in the next step of the requirements engineering process. It is such a coherent, iterative approach that is needed to draw up meaningful and value-adding requirements (Gulliksen et al., 2003).

   The lack of a coherent user requirements approach is also a problem in the e-Government context (which is the design context in this chapter). In the past, some requirements engineering activities of e-Government projects have been reported. Haraldsen, Stray, Päivärinta, & Sein (2004) discussed an approach to requirements engineering for e-Government portals that facilitates the citizen via life-events. These kinds of portals disclose all the information and services related to major events in a citizen's life, like 'getting married' or 'having a baby'. Other citizen-centric requirements studies have applied methods such as a literature review (Wimmer & Holler, 2003), a combination of interviews with experts, a literature review, surveys and focus groups (Krenner, 2002) and a combination of interviews with users and thinking-aloud sessions (Lines, Ikechi, & Hone, 2007). Although these studies report useful requirements, they do not describe a general approach for generating user requirements for e-Government services. A survey among e-Government project managers (Følstad, Jørgensen, & Krogstie, 2004) found that the key players in e-Government design need a clear and formalized approach for generating user requirements. Such an approach should include measurements that determine the success of the system design (Irani & Love, 2001).

   In this chapter, we present an approach for user requirements engineering for personalized e-Government services and illustrate it with a case study. The rest of this chapter is organized as follows: in Section 2 and 3, we discuss considerations for the requirements engineering activities in an e-Government setting and for personalization. Then, in Section 4, we present our user-centered approach and the methods involved. In Section 5 this approach is applied in a case study. Section 6, finally, rounds off this chapter with our conclusions and recommendations.

## 3.2 User requirements engineering for e-Government services

Governmental e-Services differ from their commercial counterparts. It is crucial to take these differences into account when setting up requirements engineering activities or analyzing their results. The main differences include the following:

*A heterogeneous user group*. The target group of e-Government services is highly heterogeneous as it often comprises the entire population of a region or country, while e-Commerce can focus on one single target group. Government agencies must take all the members of a population into account, which should result in a system design that caters to different cultures (Sandberg & Pan, 2007), skills (Wang, Bretschneider, & Gant, 2005), political opinions (Oostveen & Van de Besselaar, 2004), and disabilities (Becker, 2004).

*Incidental use*. Most e-Government services are used only once or rarely. As a result, clients do not have a mental model of the service they are about to apply for and must be guided through the service process by the system (Klaassen, Karreman, & Van der Geest, 2006). In the case of commercial e-Services, where the service process is more or less the same (like purchasing consumer goods), clients may have a clearer idea of the kind of service process they can expect.

*Complicated content*. Many governmental services include difficult regulations which citizens often find hard to apply to their own personal situation. In the case of e-Commerce, the service provided is usually more simple and straightforward.

*No competition*. e-Government services are usually provided by one single body and the client (citizen) is obliged to make use of each particular service (e.g., to acquire a driver's license). Therefore, e-Government services do not need to make any effort to seduce the visitor into using them as much as commercial e-Services do (Wang, Bretschneider, & Gant, 2005). As a result, there is no incentive for designers to focus on user-friendliness or attractiveness. This may result in a less usable design.

*Return on investment*. Governments use public money. This money has to be well-spent as any investments in e-Government services need to be justified afterwards (Wang, Bretschneider, & Gant, 2005). This return on investment is difficult to assess for government organizations as it often manifests itself as a reduced burden for the citizen (e.g., less effort needed to complete a service application). Therefore, Irani, Love, Elliman, Jones, & Themistocleous (2005) argue that returns on investments in e-Government

projects should be assessed using subjective user-satisfaction criteria, rather than by means of a strict economic analysis. The manner in which user requirements are formulated and the possible evaluation criteria that are attributed to these requirements can serve as the basis for such a return on investment evaluation (Jokela, 2001).

## 3.3 User requirements engineering for personalization

Not only the context (the electronic government), but also the intention to implement personalized features in the final system has consequences for the way in which user requirements engineers should work. In this section we will discuss the most important issues that need to be taken into account.

First, participants may not have any experience with personalization, and may find it difficult to envision a personalized feature if it is not made tangible for them (Weibelzahl, 2005). Consequently, it is unwise to question prospective users of the system about personalized features without any material (like screenshots or a prototypical system) that demonstrates their presence and function.

Second, if personalized features are implemented in the (prototypical) e-Service, some specific issues need to be taken into account, most notably the issues listed by Jameson (see Chapter 1). Pieterson, Ebbers and Van Dijk (2007) have discussed similar issues when pointing out potential user barriers to personalization in an e-Government context. Two closely related issues that have received a lot of attention in both the personalization and e-government literature are privacy and trust. Trust has been found to be an important antecedent of e-government service use (L. Carter & Bélanger, 2005; Horst, Kuttschreuter, & Gutteling, 2007). When the need for personalized features is discussed in the requirements, eliciting future users' perceptions of these issues should be part of the requirements validation stage.

Finally, personalized e-services can utilize data that is stored at different organizations: interoperability. e-Government services can span several organizations or departments, but can be offered to the citizen via one single website. This website should present the user with a personalized service environment that, for example, includes all relevant steps in the application process as provided by the different organizations, or utilizes information about a user that is already known at one of the associated organizations (as in pre-filled online forms). When the goal is to implement personalization in the form of interoperability in the final system, it should be assessed whether prospective users think this is value-adding and does not violate their feelings of privacy.

## 3.4 A citizen-centric requirements engineering approach for personalized e-Government services

From the array of user-centered methods that can be used in the requirements engineering stage (as discussed in Maguire (2001) and Lauesen (2002)), we selected those that suit the context of e-Government services very well. Figure 3.1 depicts our citizen-centric approach to requirements engineering for personalized e-Government services. We do not claim that this approach is the best, but will demonstrate its usefulness through the case study of the requirements engineering stage of a social support portal. This approach provides an opportunity to integrate users in the system design in a feasible and cost-effective manner. In this section we will discuss the different methods and techniques chronologically, as they are applied in the approach.
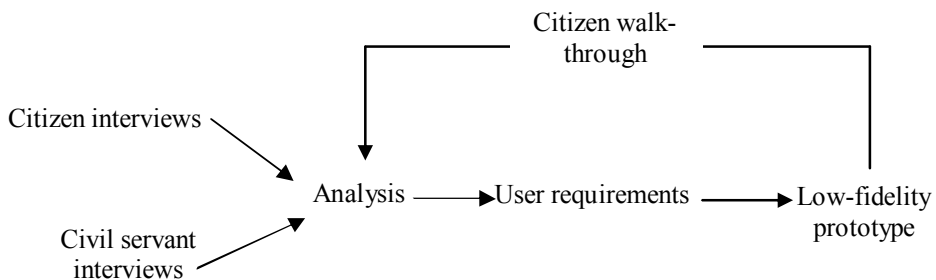


Figure 3.1. Citizen-centric requirements engineering approach

### 3.4.1 Interviews

During the interview, the interviewer asks the interviewee a series of questions about the current form of service provision and problems that arise, their goals, their expectations, and the way in which digitalization can play a role (Lauesen, 2002). The interview is also a fine method with which to identify incidents that are critical for (dis)satisfaction with the service (Gremler, 2004). Often, interviews are semi-structured: the interviewer has prepared a list of questions but is allowed to deviate from this format in order to pursue interesting issues that come up during the conversation.

For the elicitation of user requirements for e-Government services it is wise to consult stakeholders with previous and direct experience of the service in question. Two stakeholders comply most with this profile: citizens who recently applied for the service, and civil servants who are directly confronted with the service's applicants. As opposed to secondary stakeholders

who are involved from a distance, these primary stakeholders know the strong and weak points of the current, service delivery as implemented in real life, which the digital version will (partly) replace.

In requirements engineering, it is common practice to base interviews on the experienced interaction with the system to be replaced. Currently, most government services are not facilitated (completely) via the internet. In that case, the interview can provide an exploration of the context of the process which the system has to facilitate. One practical approach is to focus the discussion on the service delivery that takes place via face-to-face or telephone contact. Ambrosini and Bowman (2001) suggest encouraging citizens to tell stories about their own experiences as this makes it easier for them to speak freely and they are then more likely to share more information about the decisions made and the rationale involved.

When primary stakeholders are interviewed the focus should be on their specific knowledge. In both cases we advise using semi-structured interviews as these will enable the interviewer to pursue other potentially relevant topics that were not previously included in the interview scheme. In the case of the citizen interviews, we suggest that each conversation should address the following topics:

§ Client demographics (age, housing situation, disabilities, etc.);
§ Critical incidents that determine (dis)satisfaction with either the application process or how the application is managed, as experienced by the client;
§ The chronological service application process, as experienced by the client;
§ Expectations of digitalization of the service application and management processes.

By posing these questions, the interviewer covers the typical e-government issues of incidental use and complicated content. Furthermore, the setup allows for exploring potential users' feeling of trust and privacy when addressed during the time when expectations of digitalization of the service are discussed.

When interviewing civil servants, each conversation should include the following topics, thereby addressing the issues of incidental use, complicated content and interoperability:

§ Typical client questions or situations and their translation into actual service;
§ The information required of the client;

§ Different organizations in the service supply chain: their role, information-exchange processes and trust in the quality of information, supplied by others;

§ Expectations of digitalization of the service application and management processes.

### 3.4.2 Interview analysis

In order to generate input for the requirements formulation stage, the transcribed interviews need to be analyzed. In an overview by Davis (1982) we identified three systematic analysis techniques that are relevant for our approach. Combined, they provide the requirements engineer with an overview of the critical issues that an e-Service needs to take into account, the decisions citizens and civil servants feel they have to make and that need to be facilitated, and finally, the relevant human factors. Because each technique has its own goal, we do not think that only one of the three techniques should be selected. On the contrary, we encourage requirements engineers to apply all three techniques, as they will complement each other.

1. Critical factors analysis. This analysis technique focuses on uncovering the factors that are critical for citizens to successfully complete a process or make decisions. If addressed in the interview, the analysis can also focus on experiences which citizens deemed critical for their satisfaction with a service. This way, the requirements engineer can identify the kind of information, or the manner in which it is communicated, that is vital for an effective and efficient system.

2. Decision analysis. By analyzing the service process, as experienced by citizens, and focusing on the decisions they made, an overview of the information that needs to be provided to citizens, and at what moment, can be constructed. In order to do so, one first has to identify the (important) decisions in each process, identify the steps involved and, finally, the information that the citizen needed here. A systematic approach for this activity has been set out by McGraw and Harbison (1997). By breaking the service delivery process down into the steps as perceived by the citizen and summarizing their context, the parties involved, the information sources, and finally, the consequences of the manner in which each step is concluded, one gets a useful overview of factors and conditions that shape citizen decision making. When taking decisions made by the citizens and how they reach them into account, the system design can simplify the interaction between the citizen and the e-Service.

3. Human factors analysis. This last analysis method concerns the search for issues that may hinder successful interaction between user and sys-

tem. By taking the resulting human factors into account as user requirements in the system design, a greater fit between the system, the needs and wishes of the user, and the context can be achieved.

### 3.4.3 User requirements notation

Every critical factor, step in the decision process, or human factor that should be taken into account in the e-Service design, should be formulated as a user requirement. Several formats for the documentation of requirements are available. In order to keep this discussion clear and focused, we will discuss only one here: the Volere method[1] (Robertson & Robertson, 2006). Several features make this format superior to others in a user-centered design process.

First, the rationale behind each and every requirement needs to be written down. This will function as anecdotal evidence for the designers and, in this respect, increase the likelihood that the requirement will be implemented in the system design.

Second, and most importantly, the template forces the requirements engineer to think about the means to evaluate whether the requirement has been successfully implemented in e-Service design or not. A fit criterion specifies how the successful implementation of a requirement in (a prototypical version of) the e-Service design will be assessed, preferably by means of user evaluation. This fit criterion not only establishes the quality of the (prototypical) e-Service design, but can also determine the return on investment. When a contractor delivers a system that complies with the fit criteria, the money can be considered well-spent and vice versa. One can even draw up a contract stating that the contractor will only be paid when the system complies with the fit criteria, as in (Coble, Karat, & Kahn, 1997).

Finally, the template forces one to estimate the increase, or decrease, in customer satisfaction as a result of taking the requirement into account or not. This estimation serves as input to determine the requirements in order of priority and shows which user requirements should at least be taken into account in the final e-Service design. Writing a document that outlines the requirements marks the end of the first stage in our approach.

---

[1] The Volere requirements specification template is supported by a website which includes many resources and the template itself in different languages. It can be found at: http://www.volere.co.uk/index.htm.

### 3.4.4 Low-fidelity prototyping

Now that we have an initial set of user requirements, their relevance for stakeholders and the form in which they are to be implemented in the e-Service interface and interaction design must be evaluated. It is hard to create an instrument that makes requirements (which are often of a technical nature) understandable to prospective users (Sutcliffe, 1996). However, as we noted in Section 3.3, it is crucial that good demonstration material is available, as it allows for the evaluation of personalized features. A prototypical version of the system in combination with a real-life scenario can be used to facilitate a discussion between the requirements engineer and future users (Saiedian & Dale, 2000; Sutcliffe, 1997). Furthermore, the production of a low-fidelity prototype is a design activity that makes the design team commit itself to the formulated requirements. This way, the prospective user is ensured of a prominent role in the design process.

A low-fidelity prototype can take the form of a set of images, displaying the main screens and functionality of a system. It does not have to be representative of the final system and can be made in a program like Photoshop. Low-fidelity prototypes enable designers to quickly and inexpensively visualize the functionality and 'look and feel' of a future system, but limits the possibilities of showing the navigation within a system (Rudd, Stern, & Isensee, 1996). The use of such a prototype has been found to be a fine trigger of user feedback (Benyon-Davies, Tudhope, & Mackay, 1999) and because screenshots do not resemble a finished system in which a lot of time and effort has been invested, evaluation participants are less reluctant to provide negative feedback (Grady, 2000). Ultimately, the evaluation of a low-fidelity prototype will inform the requirements engineer whether he or she has missed some important user requirements and whether the visualized requirements are valid or not (Snyder, 2003). Examples of such prototypes can be found in Kinzie, Cohn, Julian, & Knaus (2002).

When designing the low-fidelity prototype, one should take into account the fact that it must facilitate the evaluation of the requirements for which a fit criterion has been formulated. When there are too many requirements to be assessed in a user evaluation, one will have to decide which ones will be evaluated. Such a decision must be based on the priority of the requirements (Coble, Karat, & Kahn, 1997). The more important a requirement, the more important it is that its relevance is evaluated with prospective users.

We propose a strategy that uses citizen walkthroughs, facilitated by a low-fidelity prototype and a fictive scenario. This strategy is inexpensive and easy to set up and conduct. In a system design phase where general user

feedback is needed, this approach most probably delivers the best return on investment.

### 3.4.5 Citizen walkthroughs

During a citizen walkthrough, a participant is shown the low-fidelity prototype version of the e-Service and is asked to provide comments on the functionality, the interface and the interaction design. When confronted with important functions or steps in the service process, participants can be explicitly questioned about their opinion. These questions are to be drafted before conducting the sessions and should be posed to each participant at the same time during the walkthrough. In the case of personalized e-government service, questioning participants' opinion of interoperability and trust is especially important. Traditionally, these sessions are conducted with experts, but they can be held with regular users (citizens) as well.

'Walking through' the prototype is supported by a scenario: a story about a (fictive) character that uses the e-Service. This way, the prototype functionality and its usefulness become tangible to the participant. Such characters and their stories are commonly referred to as 'personas'. Cooper (1999) describes how to choose a persona and to create a story around this character and his or her system use. When a very heterogeneous user group is to be served with the e-Service, it might be rewarding to create several personas and conduct a citizen walkthrough with representatives of the sub-groups in the end-user population, each time using the appropriate persona.

We advocate a citizen walkthrough set-up in which a low-fidelity prototype, with a limited set of screenshots (approximately 15), is presented by means of a persona. At the end of each screenshot the participant is to be asked about his or her impression of the screenshot, the completeness of the information provided, and the functionality displayed. At the end of the walkthrough, the citizen can be questioned about abstract issues such as trust, control and barriers to using the e-Service. Through this set-up, the issues of catering for a heterogeneous user group, incidental use, complicated content, privacy & trust, and interoperability are all accounted for. For a practical guide to conducting walkthroughs, we refer to Wharton, Rieman, Lewis, & Polson (1994).

### 3.4.6 Citizen walkthrough analysis

The citizen walkthroughs will result in a large amount of transcribed text. In order to generate meaningful results from these transcriptions, a systematic analysis approach is required. Based on Patton (2002), we present four analysis approaches.

1. Process analysis. This approach focuses on the user's overall perception of the e-Service process as well as the different steps contained within it.
2. Functional analysis. This approach focuses on the typical features of the e-Service, derived from the user requirements.
3. Question analysis. This approach focuses on citizens' responses to questions, related to specific screenshots or functionality, posed during the walkthrough.
4. Sensitizing concept analysis. This approach focuses on concepts that are not interface-specific, such as trust in the system or the intention to use it.

Each approach must result in the compilation of a summary of the participants' feedback about the different topics. As the citizen walkthroughs are likely to entail more than one kind of response (e.g., comments on the functionality displayed and answers to questions), more than one approach should be applied during analysis. Of course, analysis should be designed and conducted in such a way that it provides results on the fit criteria drawn up along with the user requirements.

After the citizen walkthrough, one will have to review, and possibly revise, the initial user requirements, as some will prove not to be as important as expected or will not be accepted by citizens. For all requirements, the history section of the Volere template needs to be updated. New requirements need to be added to the requirements document and, when crucial for the appearance or functioning of the e-Service, these requirements have to be tested by means of, again, a citizen walkthrough, facilitated by an updated low-fidelity prototype and scenario. When the requirements document is complete, one can start designing and programming the e-Service which, according to user-centered design principles, should also be tested with prospective users.

## 3.5 A case study in citizen-centric requirements engineering: A personalized social support e-Service

We will now illustrate our user-centered requirements engineering approach with a case study: a Dutch social support portal. First, we will discuss the application process which the e-Service is to facilitate, followed by our requirements engineering activities.

### 3.5.1 Social support in the Netherlands

Dutch citizens who, because of physical or mental ailments, cannot take care of themselves or their housekeeping, can rely on the Social Support Act

(in Dutch: Wet Maatschappelijke Ondersteuning, or WMO). If they are found to be eligible for this kind of aid, they are awarded a housekeeper or nurse for a designated amount of time and for a fixed number of hours a week. For example, John, who just underwent hip replacement surgery applies for social support as he will not be able to clean his house and has no family to help him. After being found eligible, John is given a home help for six hours a week for two months. The party that offers and arranges social support is the municipality. Municipalities often have contracts with care agencies that supply home helps and nurses, and take care of the administrative tasks involved in hiring personnel.

When applying for social support, the citizen can choose one of two options: receiving help in kind or receiving a personal budget to hire someone. When opting for help in kind, the citizen is appointed help by the municipality or care agency which then also takes on the administrative burden involved. When opting for a personal budget, the citizen can hire a home help of his or her own choice (e.g., a family member) but he or she will also have to act as the home help's employer. This means that the citizen must comply with the labor laws (e.g., maintain a system of administration to deal with salaries, etc.).

Applying for social support is difficult and involves a lot of paperwork, especially when choosing for a personal budget. In addition, each municipality has a certain degree of freedom to extend regulations regarding the Social Support Act. As the application and regulations involved differ from municipality to municipality, it is impossible to provide instructions on this process on a national level. Digitalizing the process involved in applying for a personal budget might be one way of simplifying the procedure and reducing the paperwork involved.

The requirements engineering team in our case consisted of two human–computer interaction specialists, a public administration specialist, an interface and interaction designer, and an ICT service innovator. None of us was involved in developing the final system; our task was only to deliver a set of user requirements and to advise the system developers. Exploring the use of personalized features (like a personal application process) and interoperability was included in the projects as a technology-push.

### 3.5.2 Citizen and civil servant requirements interviews

Two sets of interviews served as input for the user requirements. In the first set, we interviewed six citizens who recently completed an application for social support and consequently, could easily reflect on their experiences. This number may seem small, but at the time of interviewing the Social

Support Act had just been introduced and, therefore, the pool of applicants from which participants could be recruited was very limited. In the second set, we talked with six employees who were professionally involved in the application or administration of social support services. Here, we also spoke to six people in order to generate as much input as the applicants. The six participants represented the different professions that support social support applicants (council office clerks, application assistants and salary administration assistants).

The citizens were recruited by the municipalities that were taking part and received a gift voucher. The interviews took place in their homes, as some were physically unable to travel, and were audio-recorded. All clients had recently applied for a personal budget as a form of social support. In the end, only five of them actually received a personal budget. One citizen decided to abandon the option of the personal budget during the application process and chose to receive help in kind instead. One client was represented by a family member who also took care of the personal budget, as the client herself suffered from dementia, and one couple was interviewed together as they both received help. The employees who were interviewed were recruited through the various organizations involved in the project. These interviews were held at the offices of the employees and were also audio-recorded.

During the interviews the topics listed in Section 3.4.1 were discussed. In the interviews with the citizens, specific attention was paid to their perceptions of privacy in relation to interoperability. We wanted to know how they felt about different (semi-) government organizations exchanging personal information. Therefore, at the end of the interview, we asked citizens what they would think about different government organizations giving each other information about the interviewee, and whether there were specific kinds of information they were not allowed to exchange. In the case of civil servant interviews, we asked specific questions about their trust in the information they received from other (semi-) government organizations.

### 3.5.3 Requirements interview analysis

All audio-recorded interviews were transcribed. Next, the three analysis techniques we listed in paragraph 3.4.2 were applied to uncover requirements that were subsequently written down in the Volere template.

*Critical incidents* were only elicited from the clients. Therefore, the analysis focused on the parts of the discussion in which citizens told us about such experiences. The following example of a negative critical inci-

dent highlights the difficulty one woman had in obtaining a parking permit for a disabled person:

"I wanted to apply for it [the permit] and it caused me a whole lot of grief. In this instance, you don't get just one letter, but one for every word that's being said. I might have understood it if I had only been living in this county for two years or so… but I have been living here for 20 years. You would think that they would know me by now! I applied for it in October and received it just before Christmas. The same thing happened at the hospi-

---

**Requirement #**: 28        **Requirement Type**: Functional

**Description**: The system must provide the clients with the option of collecting data from another organization involved in the service supply chain, where the data is already known.

**Rationale**: Having to provide the same data more than once to a government agency involved in the service chain should be avoided.

**Source**: Client interview 1, 2, 3 and 5; Employee interview 1 and 3

**Fit Criterion**: Not applicable

**Customer Satisfaction**: 4        **Customer Dissatisfaction**: 4

**Priority**: High        **Conflicts**: none

**History**: Created May 1, 2007

Figure 3.2. Requirement #28 in Volere template

---

**Requirement #**: 2        **Requirement Type**: Functional

**Description**: The system must support the selection of help if the applicant opts for a personal budget.

**Rationale**: Social support clients use a personal budget to be able to determine themselves who is going to help them. The system has to support them in this.

**Source**: Client interview 1

**Fit Criterion**: Not applicable

**Customer Satisfaction**: 5        **Customer Dissatisfaction**: 5

**Priority**: High

**History**: Created May 1, 2007        **Conflicts**: none

Figure 3.3 Requirement #2 in Volere template

tal: they should just put my data into one single computer and get everything out of that."

This story, along with other statements by the interviewees, led us to formulate requirement #28 (see Figure 3.2): The system must provide clients with the option of collecting data from another organization involved in the service supply chain, where that data is already known.

*Decision analysis* was performed by means of the systematic approach by McGraw and Harbison (1997). Sometimes interviewees entrusted an agency to apply on their behalf, so in these cases a decision analysis was not performed. In the resulting tables we used the interviewees' experience of steps in the application process and choice of words as closely as possible to describe the steps they went through during the application process. An excerpt from one such table can be found in Table 3.1.

*Human factors analysis* resulted in the following statement made by a client, addressing the wish to choose a home help herself:

"I used to get help from the home help agency. But each time I got a different person. I didn't like that […] They told me I would be better off applying for a personal budget, so I could choose my own home help. Now I get the same help two times a week. A home help I chose myself."

This statement, and others like it, prompted us to formulate requirement #2 (see Figure 3.3): The system must support the applicant's prerogative to select his or her own home help if a personal budget is chosen.

### 3.5.4 Paper prototyping the social support e-Service

The citizen interviews resulted in 63 requirements while the interviews with the employees added another 39. Based upon these requirements, a low-fidelity prototype was developed which had to function as input for the citizen walkthrough. The team brainstormed about possible translations of the requirements into a visual design, which was then worked out by the interface and interaction designer. This resulted in 16 screenshots, implemented in a PowerPoint presentation.

Requirement #28 was visualized by means of collecting a citizen's net income from the Tax Administration in an e-form, and is displayed in Figure 3.4.When citizens reach a field in the e-form requesting their net income in 2006, they can click on a button 'Retrieve from Tax Service' and the displayed pop-up appears, showing the net income for 2006, as it is known by

Table 3.1 Excerpt of decision analysis table

| Step | Context | Involved parties (besides client) | Used information sources | Decision (D) or Action (A) | Problems | Simplicity | (dis)Satisfiers |
|---|---|---|---|---|---|---|---|
| ... | | | | | | | |
| Contact with Social Security Bank | Telephone contact about commissioning salary administration. | Social Security Bank | n.a. | (D) requesting form: commissioning salary administration (A) filling in form: commissioning salary administration | Client did not understand calculation of fixed monthly salary by Social Security Bank | Average | - High number of forms - Filling in the same questions more than once + Clear information |
| Renewed contact with 'support desk' at municipality | Taking care of financial affairs | 'support desk' at municipality | n.a | (A) submitting annual income form | | Average | |
| ... | | | | | | | |

Figure 3.4. Collecting net income from the Dutch Tax Administration (text in Dutch)

Figure 3.5. Selecting a home help on the basis of vicinity and appreciation by peers (text in Dutch)

the Tax Administration. Applicants can then choose to use this number in their form or to ignore it and fill it in themselves.

Requirement #2, which specifies that a client should be able to choose his or her own home help via the portal when opting for a personal budget, resulted in a separate step in the application process. This step made it possible for a client to choose a home help from a list of possible candidates. This list of candidates could be narrowed down based on their geographical proximity to the applicant, the appreciation they received from other social support clients (in the form of 1 to 5 stars), or their name. An applicant could also choose a home help on a map of his or her neighbourhood. This was visualized by means of a map of the Netherlands, containing several buckets (each representing a help). Figure 3.5 shows the screen that displays this functionality.

### 3.5.5 Citizen walkthrough of the social support e-Service

We conducted a walkthrough that was focused on the acceptance and comprehensibility of the functionality, derived from the requirements, as implemented in the low-fidelity prototype. We guided the participants through the prototype via the persona of Mrs. De Vries, who recently underwent hip replacement surgery and, via the e-Service, applied for social support, choosing for a personal budget.

The organizations involved recruited 15 participants who recently completed an application for social support, or were in the process of applying at that moment. In line with the total population that applies for social support, the percentage of senior citizens in our sample was high and included a 79-year old and an 81-year old. All in all, the participants were a representative sample of the e-Service's prospective end-users. As in the case of the citizen requirements interviews, we visited them at home and rewarded them with a gift voucher.

We would start a session by introducing the fictitious character of Mrs De Vries. Then we would show the prototype, tell Mrs De Vries' story, and, after showing each screenshot, question the participant about the visualized functionality. In the case of requirement #28 and the screen displayed in Figure 3.4, we asked each participant two specific questions, in addition to the standard questions as listed in Section 3.4.5:

§ For you personally, what are the benefits and disadvantages of retrieving data from other organizations and incorporating this directly into your own form?

§ Do you like this way of filling in a form?

After showing Figure 3.5, depicting requirement #2, we asked the participant the standard and the following specific questions:

§ Do you think it is useful that you can select a home help from a map?
§ Do you think it is useful that the appreciation of the different home helps is rated by others?
§ Do you think this is a pleasant way to select a home help?

When all the screenshots were dealt with, we posed some general questions, addressing the citizens' intention to use the e-Service, their self-efficacy, trust in the website and participating organizations, and the biggest (dis)advantage of using the website. Finally, we asked them to formulate one piece of advice for the designers of the social support portal.

### 3.5.6 Results of the citizen walkthrough

All of the citizen walkthroughs were audio-recorded, transcribed and analyzed using the approaches listed in Section 3.4.6.

The analysis of feedback on the functionality, derived from requirement #28 and depicted in Figure 3.4, resulted in the following section in our evaluation report. The suggestion to collect personal data from other organizations received an enthusiastic response. Nine interviewees told us that they thought this mechanism was pleasant and five people thought it was useful. The participants had several reasons for this opinion. The most frequently mentioned reason (four times) was that it saved them from having to search for papers. Two people stated that it prevented mistakes being made and, last but not least, two participants liked this idea because it meant that they would no longer have to do calculations themselves. One participant wondered how this would work in practice if the application was managed by a representative and thought that this alternative approach was not visualized clearly in the prototype. All in all, the results indicate that potential users greatly welcome the idea of collecting data at other organizations in order to speed up and simplify the process of filling in e-forms. Only one participant disliked the idea altogether. One point of attention should be the application process that is completed by a representative. In this case, the interface must make clear that it is the client's data that is being collected and not that of the representative.

The analysis of feedback on the functions that were derived from requirement #2 and depicted in Figure 3.5 can be summarized as follows. The functionality that facilitated the search for a home help via the e- Service received mixed feedback. Five participants were positive about the prospect of finding a home help via the website, as displayed in the low-fidelity prototype. The majority of the participants, however, provided negative feed-

back. This was caused by the spate of problems they had encountered with the functionality as displayed in the low-fidelity prototype. First of all, seven participants did not understand the function of the 'vicinity' criterion by which they could select home helps who lived close to their homes. Second, five participants did not understand the function of the stars which were used to indicate how other people rated a home help. Moreover, five of the participants indicated that they could not deduct any useful information from these ratings, as they thought they were too subjective. Third, participants told us that critical information was missing, such as someone's experience (mentioned 7 times), age (5 times) and gender (3 times). Finally, seven participants said that they would not be able to choose a home help solely via a website as they would have to meet him or her face-to-face in order to see whether they could get along well together.

### 3.5.7 Revision of the user requirements

Based on the prototype evaluation, we corrected our user requirements. In all cases, the evaluation results of the requirements that were visualized in the prototype had to be included in the 'history' section of the template.

Requirement #28 ('The system must provide the clients with the option of collecting data from another organization involved in the service supply chain, where the data is already known') was greatly appreciated by the citizens. In the Volere template a functional requirement (like requirement #28) is not awarded a fit criterion, as a function is either implemented or not. However, a positive appreciation by prospective clients can serve as a check for the right of a requirement to exist. Based on the results of the citizen walkthrough, requirement #28 should be included in the final system design in the form as it was displayed in the low-fidelity prototype, but with an added requirement, satisfying the need for clear information for representatives of social support clients, specifying how data collection goes on in their particular situation.

Requirement #2 ('The system must support the selection of a home help if the applicant opts for a personal budget') was not appreciated by participants in the form in which it was implemented in the low-fidelity prototype. This forced us to formulate additional requirements which catered for the problems that had been identified in this step of the application process. For example, a more detailed explanation about what was meant by 'vicinity' in the 'vicinity' entry field was needed, more information about the background of all the home helps was requested and finally, a function that facilitated a meeting between the two parties was to be added (a result of the need for face-to-face contact). We decided to abandon the rating feature as it

generally received negative feedback and therefore did not contribute to the usefulness of the system.

## 3.6 Conclusions and recommendations

In this chapter, we have presented a user-centered approach to requirements engineering for personalized e-Government services and demonstrated its value by means of a case study. The approach utilizes interviews, the formulation of requirements with a focus on concrete and measurable criteria, low-fidelity prototyping, and an evaluation by means of a citizen walkthrough. Based on our experiences, we will draw several conclusions on the usefulness of this approach and formulate recommendations for other requirements engineers.

Our approach, like any user-centered design process, should be seen as an iterative process: requirements, as they are translated in the prototype, need to be checked with prospective users and, if necessary, must lead to reformulated or elaborated requirements, which need to be checked again. Our case study underlined the benefits of applying more than one iteration. The need for iterative design originates in the stage in which designers develop the requirements into system design: a creative step. This interpretation may not fully correspond with users' wishes, needs or the context that prompted this requirement and thus, needs to be tested with prospective end-users. In our case study, most of the requirements, as translated in the prototype design, were accepted by the clients receiving social support. However, some requirements needed to be redefined and some additional requirements were formulated. The effort invested in the citizen walkthrough certainly proved to be worthwhile as the evaluation revealed some new issues that were crucial for successful and useful interaction between the e-Service provider and the user. The information clients need when searching for a home help, for example, appeared to be more detailed than we expected on the basis of the interviews that were geared towards determining their requirements.

We have dealt with the (possible) presence of personalized features in the final system during the user requirements engineering process in several ways. Personalization was made tangible to the participants during the citizen walkthroughs by means of a low-fidelity prototype and the scenario of Mrs. De Vries. The participants' reactions to the prototype and our questions lead us to believe that this approach has succeeded in demonstrating the personalized features. A personalized feature related to interoperability, for example, was demonstrated by automatically collecting Mrs. De Vries'

net income from the tax service. The participants provided us with well thought-out answers that showed us why they liked or disliked the feature. Based on these answers, we conclude that our approach was successful in demonstrating personalized features in low-fidelity prototypes. Next, feelings of trust and privacy were accounted for by asking interviewees to react on a scenario in which two government agencies exchanged personal information. The interviewees' comments allowed us to formulate requirements on this topic and to design a low-fidelity prototype displaying the associated functionality. The citizen walkthrough participants gave many positive comments on this feature. This indicates that the design of this instance of personalization, based on the interviews, is a successful translation of the demands and wishes expressed during these interviews, and that the method of requirements elicitation generates the comments needed to design value-adding and acceptable personalized features.

An important experience we have gained concerns the role of the requirements engineer(s). The person or persons that take on this role have a major influence on the functionality and appearance of the system that is about to be designed. Hertzum and Jacobsen (2001) have shown that experts can differ tremendously in their interpretation of evaluation results. Therefore, it is wise to put together a team of requirements engineers rather than let one single expert do all the work. Furthermore, Cooper (1999) has contended that specialists with a technical background often have dissimilar interests and beliefs on the use of technology than regular users, which may take the upper hand during the process of designing the systems. Generally speaking, this is not in the best interests of users as they might end up being burdened with functionalities they do not use or understand. Therefore, human–computer interaction specialists need to be involved. The composition of our team of requirements engineers appeared to be one way of avoiding one design viewpoint from dominating the discussion. The many debates between team members resulted in a low-fidelity prototype design that gave in to user, as well as technical, demands. As a result, the prototype not only served the wishes of the user, it was also feasible to start development. Moreover, the requirements, low-fidelity prototype and evaluation results were convincing enough for one of the largest cities in the Netherlands to take it as the basis for the development of a full-fledged, interoperable and personalized social support portal.

*Chapter 3 illustrated how one can engineer user requirements for personalization. On the basis of these requirements, a (prototypical) system can be designed. No matter how well-advanced a prototype, as soon as it makes a system's functionality clear to prospective users, it can be evaluated. As I noted before, evaluations can either be formative or summative. A formative evaluation is geared towards collecting input for redesign and is normally conducted with a prototypical version of a system. A summative evaluation is focused on assessing whether a system achieves the effects it was designed for (like increased learning performance for an e-learning system), and should make use of a well-advanced prototype or final version of the system.*

*In chapter 4, I present a literature review of publications that report user-centered evaluations of personalization. Such evaluations are focused on the subjective experience of personalization by (prospective) users. This chapter informs the reader how such evaluations are currently conducted and how this practice could be improved upon.*

# Chapter 4

## User-Centered Evaluation of Personalized Systems: A Literature Review

*"The only man who behaves sensibly is my tailor: he takes my measurements anew every time he sees me, while all the rest go on with their old measurements and expect me to fit them."*
-- George Bernard Shaw

## 4.1 Introduction

Evaluations of both personalized and non-personalized systems commonly serve three goals: verifying the quality of a product, detecting problems and supporting decisions (De Jong & Schellens, 1997). These functions make an evaluation a valuable tool for developers of all kinds of systems, because it can justify their efforts, improve upon a system or help developers to decide which version of a system to release. In the end, this may lead to higher adoption of a system, more ease of use and a more pleasant user experience.

In the literature on system evaluation, often a distinction is made between user-centered and system-centered evaluation. User-centered evaluation (UCE) is focused on gathering subjective experiences of evaluation participants. System-centered evaluation (SCE), on the other hand, aims at determining whether a system is effective and efficient or not by using a set of system metrics (Díaz, Gercía, & Gervás, 2008). Two well-known examples of such system metrics from the field of information retrieval are precision and recall. The system's performance on these criteria is determined by experts and the basis of large amounts of usage data, which does not necessarily have to originate from real-life user-system interaction. SCE is well-suited to develop efficient algorithms, or to identify the personalization techniques that score best in terms of objective effectiveness and efficiency. Besides assessing the subjective effectiveness and efficiency of a system, UCE is helpful when testing preliminary ideas, exploring a system's potential and limitations, and can generate redesign input (Petrelli, 2008). In contrast with SCE, UCE can guide the development of systems that are to be used by real people in a real-life context (Díaz, Gercía, & Gervás, 2008). Our definition of UCE is partly based on the definition of human-centered design in ISO guideline 13407: 'human-centered design processes for interactive systems' (International Organization for Standardization, 1999). We see UCE as an empirical evaluation obtained by assessing user performance and user attitudes toward a system, by gathering subjective user feedback on effectiveness and satisfaction, quality of work, support and training costs or user health and well-being.

The inclusion of personalized features causes complications for UCE. First, most traditional evaluation methods are based on the assumption that

system output is the same for each user in every context. But when personalization comes into play, this assumption no longer holds. How does one evaluate a system and generate redesign input from the results when the system constantly takes a different appearance? Second, specific usability issues, like predictability, need to be taken into account (see Section 1.4.1.). Currently, it is unclear which UCE methods are best suited for uncovering usability issues in, or assessing the perceived quality of personalization. The aforementioned complications pose us the challenge of finding valid, reliable and useful methods for UCE of personalized systems. This literature review aims to help address this challenge.

The remainder of this chapter is as follows. Section 2 will present our research question and the procedure we followed while reviewing literature. Then, we will list the kinds of personalized systems that were evaluated in the past, which variables were assessed during these evaluations and which evaluation designs they utilized, in Sections 3, 4 and 5. Section 6 describes the prototypical versions of the system that were used during the evaluation. Next, we discuss the different data gathering methods that were used and comment on their suitability for evaluating personalization. The final sections of this chapter contain points for improvement, a rough guide to evaluating personalization, implications for future research and a quickscan of evaluations published in the last few years. Section 10 ends this chapter with our concluding remarks.

## 4.2 Research question and literature selection

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

seek to inform the practitioner about avoiding the identified pitfalls and the suitability of different methods for the evaluation of personalized systems.

The following main question was formulated for the review:

*How have user-centered evaluations of personalized systems been conducted in the past?*

In order to answer this question, the following secondary questions with regard to UCE of personalization were formulated:
1. What types of personalized systems have been evaluated in the past?
2. Which variables have been assessed in the evaluations?
3. Which designs have been used for the evaluations?
4. What kinds of prototypes have been used for the evaluations?
5. Which methods have been used for the evaluations?
6. What are the advantages and disadvantages, the validity and the reliability of the methods used for the evaluations?
7. How can the usage of UCE methods be improved upon?

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

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

The following inclusion criteria were used:
§ The study had to report the evaluation of a personalized system.
§ The evaluation of the system had to be (partly) user-centered. Studies that only discussed the evaluation of algorithms or other technical aspects of the personalization process were excluded.
§ The study had to describe or discuss at least one of the following issues: advantages or disadvantages, validity, reliability, costs of a method, the participants involved, the variables assessed and the implementation of results in (re)design processes.
§ Studies reported before 1990 were not included.

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

The search was conducted from March 21, 2006 until March 28, 2006, and initially resulted in a huge number of hits (>4000). Possibly relevant hits were saved, duplicates were identified and removed, and the abstracts of 338 remaining articles were read. The selection criteria were applied to the abstracts, resulting in 127 studies of which the full text was read. The majority of these studies (n = 72) proved not to meet the selection criteria after all, and they were consequently excluded. The 55 remaining studies were included in the review. The Adaptive Hypermedia 2006 conference, which was held shortly after our search, provided six additional studies. Finally, we found two more relevant studies through references in reports (Cheverst et al., 2005; Schmidt-Belz & Posland, 2003). These were included in our review, making a total of 63 studies. A list of the articles and reports included in the literature review can be found in the Appendix.

## 4.3 Evaluated systems

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

Most of the evaluated systems were learning systems (12 studies), followed by intelligent tourist guides (8 studies), information databases (7 studies), interfaces (7 studies) and location-aware services (7 studies). The PC is the most popular platform for personalized systems, as appeared from the evaluations (42 studies). In 34 studies, the PC application was a Web browser. This does not automatically mean that most personalized systems run via a Website. Two prototype systems were designed as Websites for the sake of evaluation. Website prototypes are relatively easy to develop at low costs. Other platforms we encountered frequently were the PDA (12 studies) and the mobile phone (7 studies).

## 4.4 Assessed variables

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

1. Variables concerning attitude and experience
   § Appreciation
   § Trust and privacy issues
   § User experience
   § User satisfaction
2. Variables concerning actual use
   § Usability
   § User behavior
   § User performance
3. Variables concerning system adoption
   § Intention to use
   § Perceived usefulness
4. Variables concerning system output
   § Appropriateness of adaptation
   § Comprehensibility
   § Unobtrusiveness

In the studies in our review, 'usability' was most frequently measured, followed by 'perceived usefulness' and 'appropriateness of adaptation'. Table 4.1 shows how often each variable was addressed in the 63 reviewed studies.

Table 4.1 Variables addressed in UCE in collected studies ($n = 63$)

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

## 4.5 Evaluation designs

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

### 4.5.1 Comparing personalization and non-personalization

Comparisons aim at identifying the differences between a personalized version of a system and one in which the personalized feature is removed.

### 4.5.1.1 Usage

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

### 4.5.1.2 Implications

The validity and reliability of such comparisons have been discussed at length in the literature (e.g., Höök, 1997, 2000). Comparing a personalized system with one where the personalization has been removed is deemed a false comparison. When the personalized features have been removed, the system is no longer a worthy opponent (Höök, 2000). The evaluation reports we found also discussed this issue. Furthermore, users may be biased toward the personalized version of the system, they may favour new technology, or they may want to please the experimenter and give socially desirable answers (Bohnenberger, Jameson, Krüger, & Butz, 2002).

Finally, the study designs appeared to consider too short an interaction time to understand the full effect of personalization on the user: personalized systems need time to 'learn' about an individual user before personalization can achieve maximum performance. This implication is also known as the *cold start problem*. In order to grasp the full effect of personalization, longitudinal studies are needed (Gabrielli & Jameson, 2009; Höök, 1997).

### 4.5.1.3 Comment

Comparisons are an instrument most suited for summative SCE's. A possible outcome of such an evaluation might be that a personalized system is found to be more efficient than a non-personalized version, which may suggest that the personalized version is better. But the comparison does not tell much about usability, perceived output, or future adoption. In short, the practical value of comparisons with respect to UCD is limited (Alpert &

Vergo, 2007). Furthermore, defining quality in terms of user performance is troublesome. If it takes users longer to accomplish a task with a non-personalized version of a system, but they have a better experience using that version, one has to wonder which version is 'better'. Is there an absolute 'better' version? Evaluators must be aware of the limited importance of efficiency when it comes to the total user experience.

A comparison can be used to determine user performance with a personalized system in comparison with a non-personalized variant. However, it needs to be very clear what is being compared with what. If a non-personalized system is involved, this system should be a worthy equivalent for the personalized system and not a weak version of the original system. This might happen when the personalized features are stripped from a personalized system for comparison purposes. In that case the evaluator ends up with a system that is not optimally designed for its purpose. Comparing a personalized system and with a traditional variant (e.g., a personalized health information system and information provided on paper (see Cawsey, Jones, & Pearson, 2000)) may be fairer, and it can provide valuable insights as well.

### *4.5.2 Laboratory and real-life observations*
When observing, a researcher watches a participant working with a personalized system, noting interesting events, or recording the whole session.

## 4.5.2.1 Usage
In our collection of reports, laboratories were mentioned seven times, but none of the reports specifies how the laboratory was used. It is therefore difficult to determine whether the use of a laboratory was a plus or not. Five studies used real-life observations.

## 4.5.2.2 Implications
Using a laboratory allows the researcher to control the environment. It permits one to exclude outside influences, so one can focus on the variables one wants to assess. The downside of using a laboratory is that one loses the real-life setting; using the laboratory reduces ecological validity.

## 4.5.2.3 Comment
For some systems (e.g., a personalized website), a laboratory may facilitate a valid and reliable evaluation. The choice of an artificial or real-life testing environment depends on the relation the system has with its surroundings. If the relation with its surroundings is important to the working of a system it

might be wise to test the system in the field instead of the laboratory. An example of such a system is a personalized tourist guide that tailors its output on the basis of user interests and geographical location. It cannot be tested to its full extent in a laboratory setting.

## 4.6 Prototypes

In order to evaluate a personalized system, subjects need to interact with a final or prototypical version of the system. Otherwise, subjects have difficulty imagining how personalization will work and how it will relate to their everyday activities (Weibelzahl, 2005). The types of prototypes used in the reported studies are displayed in Table 4.2. The first three will be discussed in detail. Since not all reports specified in what stage of completion the system was during the evaluation, it was difficult to assess whether a prototype was used or not.

Table 4.2 Prototypes used for UCE in collected studies ($n = 63$)

|     | Interaction method       | Times used |
| --- | ------------------------ | ---------- |
| 1.  | Working prototype        | 17         |
| 2.  | Computer simulation      | 3          |
| 3.  | Paper prototype          | 3          |
| 4.  | Mockup simulation        | 2          |
| 5.  | Wizard-of-Oz prototype   | 2          |
| 6.  | Demonstration of system  | 2          |

The use of a working system prototype was reported 17 times. Creating a working prototype is relatively easy and cheap compared with creating a full system, and it can help to verify or improve upon the quality of a product. Hence, creating a prototype is a wise use of money and effort (Field, Hartel, & Mooij, 2001). However, the use of a prototype version of a system does not yield the same results as one would obtain using the full system (Field, Hartel, & Mooij, 2001) and a prototype that has been simplified too much may not be the best instrument for uncovering usability issues.

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

Low-fidelity prototypes, like paper prototypes, allow the researcher to test personalization in a very early phase of system development (as in Karat, Brodie, Karat, Vergo, & Alpert, 2003). They can help to pinpoint crucial issues that may play a role in the adoption of the system in the future and can help to assess the participants' attitudes towards personalization (Walker, Takayama, & Landay, 2002). Paper prototypes often include just a few screenshots of the system, which are not necessarily representative of the final version (see, for example, the screenshots we used in Section 3.5.4). They must provide the participant with an idea of the system functions and personalized output (Virzi, Sokolov, & Karis, 1996). Based on the low-fidelity prototype, participants can comment on the concept behind the system and its functionality. Because of their abstract nature, these prototypes are likely to elicit mainly abstract user feedback.

Table 4.3 Methods used for UCE in collected studies ($n = 63$)

| | Method | Times applied | Variables most often assessed |
|---|---|---|---|
| 1. | Questionnaire* | 47 | Usability (21) |
| | | | Perceived usefulness (19) |
| | | | Intention to use (14) |
| 2. | Interview* | 27 | Intention to use (9) |
| | | | Appropriateness of adaptation (7) |
| | | | Usability (7) |
| 3. | Data logging | 24 | User behavior (17) |
| | | | User performance (6) |
| 4. | Focus groups & group discussions* | 8 | Intention to use (4) |
| | | | Perceived usefulness (2) |
| | | | Trust & privacy issues (2) |
| 5. | Thinking-aloud* | 7 | Usability (2) |
| | | | User behavior (2) |
| 6. | Expert review | 6 | Usability (3) |

Note: Only methods that are used more than twice are shown
*UCE method

## 4.7 Data gathering methods

The methods and instruments used in the investigated studies are listed in Table 4.3. Two points are worth noting. First, almost every study uses multiple data gathering methods. Second, some of the listed methods (e.g., data log analysis) may not be in accordance with our definition of UCE, because they do not collect subjective feedback from (potential) users. However, studies that used methods that are not user-centered along with methods that are user-centered were still included in our review. In these cases, the methods should be seen as complementary, providing triangulated data, and con-

tributing to the overall value of the evaluation. In the following subsections the methods listed in Table 4.3 will be discussed more thoroughly.

### 4.7.1 Questionnaires

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

## 4.7.1.1 Method usage

Questionnaires were the most common evaluation method in the studies, and were used 47 times. The most commonly measured variables were usability (21 times), perceived usefulness (19 times) and intention to use (14 times). Of the questionnaires, 30 only presented closed questions, 5 only open questions, 9 both kinds of questions, and for three studies it could not be assessed what kind of questions were posed. Interestingly, questionnaires were often used both for gathering global impressions of the system and as a tool to identify problem areas. In 25 studies, the questionnaire was used as a form of formative evaluation and in 16 studies for summative evaluation.

One of the advantages of questionnaires is the large number of participants that can be accommodated. Table 4.4 shows the different sample sizes used in the reviewed studies. The number of questionnaires completed by only a small number of subjects is large and questionable from a methodological perspective. If a sample group is small, it is difficult to generalize findings (Weibelzahl, Lippitsch, & Weber, 2002) and a small number of respondents limits the possibilities of statistical processing. Some authors discussed the need for a large group of respondents in order to generate a meaningful evaluation (Cawsey, Jones, & Pearson, 2000) or commented that their study suffered from the small number of respondents (Gregor, Dickinson, Macaffer, & Andreasen, 2003; R. Henderson, Rickwood, & Roberts, 1998). In order to make claims about the general usability of a system, for example, sample sizes need to be large enough that they can be generalized to a larger population of users (Dicks, 2002). In the evaluation design, the designated number of participants should be geared to the goals of the evaluation. Whereas problem detection can be done with a relatively small sample, the verification of quality needs a larger, representative group of respondents.

Table 4.4. Number of questionnaire respondents

| Number of respondents | Times encountered |
|---|---|
| <25 | 15 |
| 25-100 | 23 |
| 100-250 | 6 |
| >250 | 4 |

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

Measuring psychological constructs with only one or two items (e.g., R. Henderson, Rickwood, & Roberts, 1998) is generally not considered good practice, and measuring two concepts in one question might easily yield non-valid results (Spector, 1992). An example of the latter was found in Stary & Totter (2003). They asked participants: 'Overall, how clear and useful are the contents of the information system?' Clarity and usefulness are two distinct variables, but here the participants must express them in one answer on a Likert scale. As a result, no definite answer can be expected.

Finally, the design of a Likert scale should result in a measurement continuum (Spector, 1992). Muntean and McManis (2006), for example, designed a five point Likert scale to assess usability and user satisfaction. Their scale included the response choices: 1, poor; 2, fair; 3, average; 4, good and 5, excellent. The inclusion of 'fair' as the number 2 choice distorts the continuum of the Likert scale. Therefore, one cannot interpret an average score on one of these items as a valid indication of usability or user satisfaction: the results are skewed toward a neutral or positive reaction.

The results of questionnaires are not reported systematically and as a result, it may be hard for a reader to judge the quality of an evaluation. When assessing quantitative scores on variables by means of multiple items and Likert scales, it is good practice to first determine and report the quality of measurement. By determining and reporting Cronbach's $\alpha$ (Cronbach, 1951) for each variable, readers can assess the quality of a questionnaire. Next, reporting the means and standard deviations for each variable can provide the reader with a clear view on the answers the participants gave and their consensus (or lack thereof) on a certain topic.

## 4.7.1.2 Comment

A questionnaire can be very useful for measuring appreciation of personalization, user satisfaction, general opinions about the system, or for benchmarking purposes. Such variables are suitable for quantitative measurement, using Likert scales, for example. Questionnaires, with either open or closed questions, may not be the best choice for identifying usability problems. Other methods that collect their data simultaneously with the user-computer interaction may be suited better as they can capture problematic incidents 'on the fly'. The next chapter will report a study that delves into this matter.

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

### 4.7.2 Interviews

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

## 4.7.2.1 Method usage

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

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

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

## 4.7.2.2 Comment

The interview is a fine method for exploring abstract issues in depth. Such issues include the specific usability issues for personalization as listed in Section 1.4.1, appropriateness of personalization and the appreciation of personalization.

The reports of interviews in the articles were generally not very detailed. The interviewer and his background often remain unmentioned and systematic overviews of the data collection process and results are scarce. In reports, the processing of data from the interviews has to be described and justified, in order to determine its quality. For an overview on how to ana-

lyze qualitative data, like interview results, we refer to Miles & Huberman (1994). When reporting interviews, a good standard is to provide typical comments made by interviewees (e.g., Ketamo, 2003; H. Smith, Fitzpatrick, & Rogers, 2004).

### 4.7.3 Data log collection

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

## 4.7.3.1 Method usage

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

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

## 4.7.3.2 Practical implications

Analyzing data logs to determine user performance may be troublesome. Stein (1997) reports that it may be difficult to interpret task completion time as a measure of user performance (the faster a subject completes a task, the more efficient a system), as personalized output may sometimes provide the user with more (difficult) output on purpose (e.g., when the system assumes a student needs more reading material to understand a lesson).

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

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

### 4.7.3.3 Comment

We strongly advise using data log analysis to assess user behavior or user performance in combination with a qualitative method. The use of triangulation provides deeper understanding of results generated with data log analysis (Ammenwerth, Iller, & Mansmann, 2003).

### *4.7.4 Focus groups*

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

### 4.7.4.1 Method usage

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

The reporting of focus group results lacks the same detail as interviews (see Section 4.7.2). Evaluators do not systematically report the answers provided by the participants, but present trends. Hyldegaard & Seiden (2004), for example, state that 'site information, actuality, language and layout of the interface were often used by the participants as arguments for trusting or not trusting a personalized output'. From such a statement, we cannot assess whether any of these arguments were mentioned more often and therefore, are perhaps more important.

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

### 4.7.4.2 Comment

Focus groups are suitable for gathering a large amount of qualitative data in a short time. They may help to provide a deeper understanding of abstract issues like user behavior, user appreciation of personalization, user perceptions regarding the specific usability issues for personalization, and the intention to use a personalized system. Another interesting and promising ap-

proach is to combine focus groups and paper prototyping. This can generate useful design input in an early phase of the development process (Kaasinen, 2003; Karat, Brodie, Karat, Vergo, & Alpert, 2003).

### 4.7.5 Thinking-aloud

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

## 4.7.5.1 Method usage

Seven studies used think-aloud protocols. The quality of reporting of think-aloud protocols was poor. Researchers often assume that mentioning the use of such protocols is enough for the reader to understand what is being done. It would be a positive addition to evaluation reports if writers would mention to what goal thinking-aloud was used.

Thinking aloud often happens in congruence with test tasks. These test tasks determine the scope of the user–system interaction during the evaluation and therefore, can influence results. In order to guarantee a valid evaluation that takes into account all the main aspects of the personalized system, the test tasks need to address them. When presenting the evaluation methodology, test tasks need to be clarified and show a well-balanced distribution of system functions over test tasks. Statements like 'Visitors were asked to perform eight tasks in total. They were also asked to think aloud while interacting with the system' (Pateli, Giaglis, & Spinellis, 2005, p. 203) do not suffice.

## 4.7.5.2 Comment

Think-aloud protocols can provide fruitful insights for understanding the causes that drive user behavior and can be a great source for identifying usability problems.

### 4.7.6 Expert reviews

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

### 4.7.6.1 Method usage

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

## 4.7.6.2 Comment

Although the method is cost-effective, one can doubt whether expert reviews are truly the best way to identify usability problems with (personalized systems). Literature indicates that experts are not capable of supplying the same critique on a system that (potential) end-users can. Van der Geest (2004) found that expert reviews uncover different issues than focus groups or think-aloud protocols (both of which involve end-users). Lentz & De Jong (1997) showed that experts are unable to predict user problems because experts have different skills, knowledge, cultural backgrounds and interests than users. Hartson et al. (2001) agree, stating that the realness of usability problems identified by experts can only be determined by real users. Finally, a study by Savage (1996) identified a discrepancy between expert and user reviews. Although experts identified areas of an interface that needed to be further investigated, users focused their comments on changes that had to be made to the interface.

A method of evaluation that is closely connected to expert review is heuristic evaluation. Heuristics are design rules for a system, and in a heuristic evaluation an expert determines whether a system has been designed in accordance with these rules. Expert reviews and heuristic evaluation are in essence both applications of a set of design rules (either implicitly stored in an expert's memory or explicitly written down) to a system. Kjeldskov et al. (2005) showed that applying textbook heuristics is too static to assess the usability of personalized systems. It is an inappropriate method, as it is unable to account for different user characteristics and contexts, and thus uncovers only a limited number of usability problems. Experts may 'suffer' from the same limitation, being unable to identify with every user and with every context the personalized system is designed for. However, Magoulas, Chen and Papanikolaou (2003) showed that heuristics, altered to fit the system, can be useful when formulating redesign guidelines. And Welle-Donker Kuijer, De Jong and Lentz (2008) state that heuristics might be beneficial for expert evaluations when they have novelty value for the experts. Since evaluating personalization might be a new experience for experts, heuristics may be a useful tool for expert evaluation. This discussion has made it clear, as we will argue in Section 4.8.3, that the added value of heuristic and expert evaluation for evaluating personalization needs to be a topic of future research.

There are possible drawbacks in using expert reviews or heuristics to evaluate personalization. Evaluations conducted with (potential) end-users might yield more valuable results, as they can account for the different us-

ers' characteristics and contexts involved. On the other hand, expert review or heuristics may be valuable tools for evaluating specific topics not involving personalization, like accessibility and legal issues.

## 4.8 Conclusions

### 4.8.1 Improving UCE of personalization

When we reflect on the methods which have been applied during evaluations, we see that questionnaires are very popular. Using questionnaires during evaluation makes sense during summative evaluations, but the method is not so popular during user-centered design in general (Vredenburg, Mao, Smith, & Carey, 2002). Some of the appended questionnaires we encountered in our review were poorly designed, which might be cause for concern. Questionnaires seem to be used as the quick and dirty way of assessing user opinion about personalized systems. As a result, the potential of questionnaires is not fully exploited, and the quality of the resulting evaluations is low. The reason for using a questionnaire needs to be considered carefully. Issues that call for an extensive review of the system, like usability, require other, more exhaustive methods. Questionnaires with closed items can only confirm known variables and assess their scores. Qualitative methods, such as interviews, focus groups or thinking, allow researchers to explore issues (like system trust or controllability) in depth, or make it possible to identify unforeseen problems. As a result, they are more valuable to the system design process.

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

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

Finally, there are some variables whose importance we would like to stress and which can be assessed for the majority of personalized systems, namely the specific usability issues for personalization. The reports we read did not generally assess these usability issues (predictability, controllability, unobtrusiveness, privacy and breadth of experience), as compiled by Jameson (2003; 2007; 2009). Because these issues determine the shape of the user experience, they can prevent the user from wanting to interact with a personalized system (e.g., if the user perceives the collection of user data as a serious infringement on privacy). Therefore, dealing with theses specific usability issues is an important part of a UCE of a personalized system.

### 4.8.2 A rough guide to UCE of personalization

In this section, we will discuss the use of different UCE methods during the UCD process for personalization. During the UCD process, one can use four different kinds of (prototypical) systems. Each (prototypical) version of the system can serve one of the goals of evaluation (verifying the quality of a product, detecting problems and supporting decisions (De Jong & Schellens, 1997)) best. A similar overview per method has been published by Gena & Weibelzahl (2007).

At first, there is no system or prototype at hand, so studies can only assess the context in which the system is supposed to function, or gather features the end-user would like to see in it. Such a situation is mostly present in Maguire's (2001) context of use and requirements phases. The use of UCD methods in a situation where there is no (prototypical) system is outside the scope of this chapter.

Next, one can make use of low-fidelity prototypes in the UCD process. Evaluations of such prototypes deal with the ideas behind the personalized system and the techniques that are supposed to embody these ideas. A low-fidelity prototype can visualize them, and from this moment on, UCE becomes an option. The methods that are most appropriate here are those that can collect in-depth data. Qualitative methods, like interviews or focus groups, can be used to assess perceived usefulness, future system adoption, appreciation, trust in the system, etc.

Once the actual programming is under way, a working prototype can support evaluations that address both usability and the usefulness of personalization in individual settings. Qualitative methods, like thinking-aloud, interviews or observations will be particularly suited to identify general usability issues and the specific usability issues for personalization. Quantitative methods like questionnaires with Likert scales can be used here for benchmarking purposes (e.g., to assess user performance), although it

should be kept in mind that a prototypical version of a personalized system may not be a suitable reference point for a benchmark.

Finally, one has the possibility of using a finished system for evaluation purposes. The transition from formative (prototype) to summative (full system) evaluation means a change in evaluation goals. Where a prototype evaluation mostly focuses on problem-oriented results, a full system evaluation will focus on benchmark results. Here, quantitative methods like questionnaires with Likert scales will be best suited to assess variables like user satisfaction, user experience, or perceived user performance. Qualitative measures like thinking-aloud and interviews may be very valuable to explain the results obtained with quantitative methods. However, one should be cautious to combine testing for efficiency with thinking-aloud, as the latter method may have a negative influence on task performance (Van den Haak, De Jong, & Schellens, 2003).

### 4.8.3 Implications for future research

In this section, we will outline, what we think are the most crucial issues that need to be taken into account in future research or evaluations.

At the moment of writing, the pros and cons of different UCE methods for evaluating personalization have not been fully documented. It is crucial that we know how each method can take into account the specific usability issues for personalization and can elicit user perceptions on the usefulness of personalization in the context of use (Akoumianakis, Grammenos, & Stephanidis, 2001). It would therefore be interesting to have 'showdowns' between UCE methods evaluating the same personalized system. These comparisons between methods would allow us to see which method is suited to elicit a specific kind of participant feedback. Such knowledge will allow us to create better evaluation designs for personalized systems. However, comparing methods has a possible limitation as we must be cautious about generalizing their results. Results that are found for one kind of personalization do not necessarily hold for other kinds of personalization as well. Furthermore, some types of personalized systems provide a very different appearance to different users (like personalized systems that convey information via different modalities, suited to the disability of a specific person). In these cases, one can wonder whether or not it may be necessary to assess a method's validity, reliability, advantages and disadvantages for every appearance of the system.

The next chapter reports on a comparison between the usefulness of thinking-aloud, questionnaires and interviews for the formative evaluation of a personalized search engine. Another interesting comparison that future

research needs to make is between expert reviews or heuristic evaluation on the one side, and user evaluation on the other side. Expert reviews or heuristic evaluations are very cost-effective means for identifying usability issues in non-personalized systems. But does this premise also hold when every individual is presented with tailored system output, and usability issues and perceptions of usefulness are highly dependent on the unique individual context? Or is consulting (prospective) users a more fruitful alternative?

## 4.9 Quickscan: 2006-2010

The collection of publications for this literature review was conducted in March, 2006. As a result, the results may be somewhat outdated, so we conducted a quickscan in September, 2010 to see if new topics in UCE of personalized systems emerged. This was done by surveying the issues of the journal User modeling and user-adapted interaction from issue 1 in 2006 to issue 3 in 2010, and the conference proceedings of User modeling 2007, Adaptive Hypermedia 2008, and User Modeling, Adaptation, and Personalization 2009 and 2010. We haven chosen this journal and these conferences as the collection of reports in 2006 pointed out they are the premier sources for publications that include evaluations of personalized systems.

The quickscan identified 37 UCE reports. The collected evaluations include, as was the case in the initial collection of reports, a substantial amount of summative evaluations that consist of comparisons (participants interacting with a personalized system or a version from which the personalization features have been removed) followed by questionnaires and backed-up with data logging results (e.g., Niu & Kay, 2008). There does not seem to be a significant shift in method usage. Also similar to the first data collection are the kinds of systems that are evaluated: adaptive learning systems, museum guides and location-based tourist guides make up the majority. On the other hand, we have identified three interesting developments that seem to be taking place.

First, there seems to be a growing interest in *user acceptance* of personalization and this is reflected in the high number of studies delving into this matter. Cramer et al. (2008), for example, utilized questionnaires, interviews and thinking-aloud to gain insight into users' decision whether or not to use an adaptive art recommender and took into account the role of perceived system competence and trust in the system. Other researchers only made use of questionnaires to determine, for example, user acceptance of an adaptive museum guide (Pianesi, Graziola, & Zancanaro, 2007) or recommender systems that use personality quizzes (Hu & Pu, 2010).

Second, *trust* (often defined as a person's confidence that a system will provide relevant and high-quality output) is an issue that appears to be receiving more and more attention in evaluations of personalization. For example, Bunt, McGrenere and Conati (2007) studied trust in the context of customizable user interfaces and Tintarev and Masthoff (2008) studied the issue with respect to providing movie recommendations with or without adaptive explanations.

Third, in contrast to the publications before March 2006, in the years 2006 to 2010 more reports of *iterative design processes* seem to have been published that include several rounds of UCE. Carmagnola et al. (2008) report on the development of an adaptive tourist guide in which they used heuristic evaluations, usability tests and interviews. The problems that were identified in each evaluation were consequently addressed in a redesign of the system. Zimmermann & Lorenz (2008) report the design process of a personalized museum guide. They made use of expert reviews (two iterations) and user evaluations with questionnaires (open and closed questions) and interviews. The authors reflect on the use of the expert reviews, stating that these evaluations "give an idea of what the benefit brought along by personalization might or might not be, both for the museums and their visitors." (Zimmermann & Lorenz, 2008, p. 413).

The quickscan suggests that UCE is gaining popularity in the personalization community and the developments we have discussed above indicate that the prospective user is more and more a central focus during the design process. If this trend continues, personalized systems are likely to be more usable and chances are that they will be used by a higher number of people.

## 4.10 Closing remarks

UCE can be of great value in the UCD process of personalized systems as it can provide a sound basis for improving the interaction between user and system. At this moment, we do not know a great deal about the suitability of different evaluation methods for evaluating personalization, due to the relatively early stage of research into personalized systems. This review has mapped the current practice of UCE of personalization. We hope that it can serve as a starting point for improving UCE's of personalization. It is crucial that evaluators create well-considered evaluations, by using the right design, prototype and data gathering method(s), and afterwards, report their procedure and findings in a complete and understandable manner. We are convinced that improving the UCE practice of personalization will lead to better personalized systems and will ultimately serve the user.

*In the previous chapter, I have mapped the current user-centered evaluation practice of personalization. This practice needs to be improved upon. In addition, we do not know the yield of the different methods when applied during a user-centered formative evaluation of personalization.*

*Chapter 5 explores the value of thinking-aloud, questionnaires and interviews for the formative evaluation of a personalized system. It focuses especially on each method's ability to elicit user comments on the specific usability issues for personalization, appreciation of personalization and finally, comments on the quality of personalization. The results of this study allow evaluators to make better informed choices when they have to decide upon a suitable method when conducting a user-centered formative evaluation of a personalized system.*

# Chapter 5

## Identifying Usability Issues for Personalization during Formative Evaluations:
## A Comparison of Three Methods

*"A common mistake people make when trying to design something com-*
*pletely foolproof is to underestimate the ingenuity of complete fools."*
-- Douglas Adams

## 5.1 Introduction

With the introduction of personalization, evaluators have to ask themselves several questions. Are the evaluation methods we have always used during formative evaluations suitable for eliciting user comments on the specific usability issues for personalization? Which method or combination of methods can we best use to identify problematic issues with the personalization provided to users? And how do we elicit user comments on the quality of personalization?

As we have seen in the previous chapter, currently we do not have a full grasp of the ability of the different evaluation methods for dealing with personalization. As a result, we recommended conducting showdowns between the different methods for assessing this ability. As we have seen, three popular methods for evaluating personalization are thinking-aloud, questionnaires and interviews. In this chapter, we will report on a comparison between these methods for evaluating personalization during formative evaluations.

Formative evaluations focus on identifying the largest number of problems with a system or website. These problems should consequently be solved by redesigning the system. When launching a personalized system or website, one will want to have solved any problems linked to these specific issues. By doing so, a high degree of usability and a pleasant user experience can be achieved. The results of this study will aid evaluators in the decision on choosing the best method for eliciting user comments that can serve as input for personalized system redesign.

In this chapter, we will first introduce the three methods that will be compared. Next, Section 3 will outline our research question and hypotheses, followed by a discussion of the study setup in Section 4. Results can be found in Section 5. We will end this chapter with our conclusions and a discussion.

## 5.2 Thinking-aloud, questionnaires and interviews

As the previous chapter showed, three methods appear to be popular for evaluating personalization: thinking-aloud, questionnaires and interviews.

### 5.2.1 Thinking-aloud

Thinking-aloud is a method that draws out participants' inner thoughts or cognitive processes while they are engaged in interacting with a system (Patton, 2002; Peleg, Shackak, Wang, & Karnieli, 2009) and encourages them to reflect on their own behavior (Van Oostendorp & De Mul, 1999). It can be used to identify unsatisfactory features of a website (Benbunan-Fich, 2001) and reveals the usability problems that users encounter when they are busy interacting with a system (Jaspers, 2009), as well as general comments about a system (Hoppmann, 2009). Gena and Weibelzahl (2007) claim that, for personalized systems, thinking-aloud can be conducted to elicit comments on users' cognitions and their thoughts on the usability of interface adaptations.

### 5.2.2 Questionnaires

Questionnaires may include two different kinds of questions: closed or open-ended. Closed questions (e.g., statements with scoring scales) can pinpoint problem areas or can be suitable for benchmarking purposes. For example, they can help to compute a score for the comprehensibility of a prototype and the final version of a system. These scores can then be compared to determine whether the changes made have affected users' comprehension of the system. However, these scores will not tell us anything about *why* a user does or does not comprehend a system (Kushniruk & Patel, 2004) and that is invaluable information when one wants to improve the system. According to Labaw (1981), closed questions have another caveat: they do not give any indication of whether or not the participant actually understood the topic under investigation or if he or she is simply being conscientious about filling in all the questions. Open-ended questions can provide the information that closed questions do not give (R. D. Henderson, Smith, Podd, & Varela-Alvarez, 1995; Miles & Huberman, 1994). They offer the participant the opportunity to explain the rationale that informs their opinion about a psychological construct (Bradburn, Sudman, & Wansink, 2004). A downside of the questionnaire is that the scope of the participants' answers is limited to the subjects covered by the questionnaire (P. Carter, 2007). Gena and Weibelzahl (2007) claim that questionnaires can inform the evaluator about participants' opinions and satisfaction rates regarding a personalized system. Such questionnaires are commonly given to participants' after they have interacted with the system that is under evaluation (Kaufman, 2006).

### 5.2.3 Interviews

Interviews may be structured or semi-structured. In a structured interview, the interviewer is obliged to follow the interview guidelines and cannot probe more deeply into any unexpected issue that crops up during the conversation. However, in a semi-structured interview, the interviewer is allowed to do this. Like open-ended survey questions, interviews can supply the evaluator with feedback on a given, general topic (Fossey, Harvey, McDermott, & Davidson, 2002). A downside of semi-structured interviews lies in the freedom an interviewer enjoys regarding the sequence and wording of questions. This may influence responses which makes it hard to compare comments on a given topic (Patton, 2002). Gena and Weibelzahl (2007) claim that, for the case of personalization, interviews are the most effective method for assessing user opinions and satisfaction levels.

### 5.2.4 Comparisons of methods

A considerable number of comparisons between usability evaluation methods address the differences between expert review methods and user methods (Doubleday, Ryan, Springett, & Sutcliffe, 1997; Jeffries, Miller, Wharton, & Uyeda, 1991; Lentz & De Jong, 1997; Savage, 1996). Other comparisons have addressed the differences between thinking-aloud, questionnaires and/or interviews when applied to tasks or systems without personalized features. In the case of text-processing, Scott (2008) found that thinking-aloud and interviews elicit the same responses. Meanwhile, other researchers did find differences between the methods. In a comparison conducted with child participants, Donker and Markopoulos (2002) found that thinking-aloud uncovers more usability problems in an educational game than questionnaires and interviews. Furthermore, these last two methods did not differ significantly in the number of problems they uncovered. After comparing different evaluation methods, Ebling and John (2000) concluded that a combination of performance and questionnaire data will uncover the most critical problems, while thinking-aloud will give the evaluator the largest overview of all usability problems. Henderson et al. (1995) also arrived at the conclusion that thinking-aloud identifies the largest number of usability problems, when compared to interviews, questionnaires or data log analysis. They also advise evaluators to use questionnaires with open-ended questions in order to generate the most useful feedback. According to Allwood and Kalén (1997), thinking-aloud elicits the most comments from text readers and identified most problems when compared to participants underlining problematic text parts or writing down questions.

In order to reveal the full spectrum of a system's strong and weak points, one needs to evaluate it using multiple methods (Ebling & John, 2000; Peleg, Shackak, Wang, & Karnieli, 2009; Scott, 2008; Zabed Ahmed, 2008). However, the sets of issues the different methods elicit may overlap and, as a result, the added value of applying an extra method may be limited. Henderson et al, for example, found the added value of using other methods in combination with thinking-aloud to be limited (R. D. Henderson, Smith, Podd, & Varela-Alvarez, 1995). Therefore, during a comparison, it is important to assess the relative merit of the various user-centered evaluation methods.

## 5.3 Research question and hypotheses

Our research question addresses the knowledge gap concerning the ability of three user-centered evaluation methods to elicit participants' comments on specific usability issues for personalization and the perceived quality of personalized output. We will compare the comments on personalization elicited through thinking-aloud (a method applied during the actual process of interacting with the system) on the one hand, and questionnaires and interviews (which are methods applied after interaction with the system has taken place) on the other. Hence, our research question is:

*What is the yield of thinking-aloud, questionnaires or interviews when applied during the formative evaluation of a personalized system?*

When evaluating a personalized system, one can collect comments on both the issues that are specific for personalization, as well as generic issues. Generic issues are issues that are neither specific for personalized systems nor influenced by personalization. In other words, they are the usability issues identified in a personalized system that are unrelated to personalization. 'Receiving unexpected search results from a personalized search engine', for example, is a specific issue, while we would consider the 'misunderstanding of the working of a drop-down menu' to be a generic issue.

We will test our hypotheses using one specific form of personalization: personalized link sorting. According to Knutov, De Bra and Pechenizkiy (2009), this is a personalized presentation technique. Link sorting is concerned with the generation of a list of links, ranked according to user characteristics and interests. This technique is applied in a large number of different systems or websites such as personalized search engines or personalized e-learning systems. We have chosen this form of personalization as it is very

salient: users will most probably notice that output is being tailored. As a result, we are more likely to receive feedback on personalization than if we had used another personalization technique. The personalized hiding of irrelevant links on a website, for example, may go unnoticed by users because they only see the links that are there. They may very well be unaware of the fact that something is being hidden.

### 5.3.1 Specific issues

Our first hypothesis addresses the ability of three user-centered evaluation methods to elicit comments on the specific usability issues for personalization (predictability, comprehensibility, controllability, unobtrusiveness, privacy, breadth of experience, system competence) and one related issue: appreciation of personalization. As no studies have delved into this matter before, our hypothesis (H) is that the three methods perform equally well.

H1. Thinking-aloud, questionnaires and interviews yield the same number of comments from participants on specific usability issues and appreciation of personalization.

The success of personalization should be assessed by focusing on its main objective (Weibelzahl, 2005). Only then can the usefulness of the output for a specific user in his or her context be determined. Since we compare, in this study, the different methods using the case of a personalized search engine, usefulness can be interpreted as the *perceived relevance of search results* (Nahl, 1998). Participants are constantly confronted with search results while interacting with the system. As thinking-aloud is a method that draws out participants' inner thoughts during interaction, it is most likely that this method will perform best at gathering comments on perceived usefulness. It is 'closest to the fire'. Based on Carroll et al. (2002), thinking-aloud is hypothesized to be the best method for obtaining the inner thoughts that precede this judgment of usefulness.

H2. Thinking-aloud elicits more comments from participants on the perceived relevance of personalized search results than questionnaires and interviews.

The third hypothesis deals with the value (positive, negative or neutral) of each comment on personalization. Comments on personalization are the collection of comments on usability issues for personalized systems, the appreciation of personalization, and comments on the perceived relevance of

search results. Thinking-aloud is superior to questionnaires and interviews in identifying unsatisfactory features (Benbunan-Fich, 2001). An unsatisfactory feature, which may need to be improved upon during the redesign process, will result in negative comments about the system (output). Thus, thinking-aloud will elicit more negative comments than the other two methods.

H3. Thinking-aloud elicits more negative comments on personalization than questionnaires and interviews.

Finally, thinking-aloud is assumed to elicit more interface and interaction-specific comments on a system (Benbunan-Fich, 2001; Patton, 2002). Questionnaires and interviews are more effective for eliciting statements on general topics from participants (Fossey, Harvey, McDermott, & Davidson, 2002). So, in accordance with a study by Ebling and John (2000), the set of issues identified by thinking-aloud on the one hand, and questionnaires and interviews on the other, should differ.

H4. The problems related to personalization identified by thinking-aloud on the one hand, and questionnaires and interviews on the other, do not overlap.

### 5.3.2 Generic issues

Of course, a formative evaluation of a personalized system needs to uncover a lot more than just issues related to personalization. Generic issues can have a detrimental effect on the usability and usefulness of a system and, therefore, need to be dealt with during the redesign process. The ability of formative evaluation methods to uncover generic issues has been reported in the past (see Section 5.2). As the goal of this study is not to compare the success of the three methods in eliciting comments on generic issues, we will treat the results concerning generic issues generally. As a result, we did not formulate any hypotheses for these analyses.

## 5.4 Method

After explaining the system we evaluated in this study, we will describe our experimental procedure and analysis of the collected data. Finally, we will discuss how we avoided the pitfalls that are part and parcel of evaluating a personalized system.

### 5.4.1 Research context: Prospector

We tested our hypotheses using a personalized search engine called Prospector (Schwendtner, König, & Paramythis, 2006) which applies personalized link sorting. We chose this system for four reasons. First, at that moment, the system was still in development so we could expect that at least some problematic issues could be identified. Second, the link sorting done by Prospector is explicit. Participants will notice that they are interacting with a personalized system. Third, as we will explain below, Prospector's user profile can be viewed and altered by the users. This feature, which allows participants to give detailed comments about the quality of the user profile created by the system, is becoming increasingly popular in personalized system design. Moreover, it contributes to system controllability, which chapter 2 has shown to be a very important aspect of personalized system design. Fourth, Prospector is a search engine and participants will therefore have a point of reference for their judgment of quality: Google. This can make it easier for them to comment on the quality of Prospector.

Prospector was developed by the Institute for Information Processing and Microprocessor Technology at the Johannes Kepler University in Linz, Austria. It is an internet meta-search engine that re-ranks search results from primary search engines (such as Google or Yahoo) on the basis of a user model consisting of user interests and user ratings from previously visited search results.

When using Prospector for the first time, users indicate their interests in 13 general categories (e.g., art, sports, etc.). During the searches (see Figure 5.1), user ratings of search results are collected via a rating frame that is displayed above each opened search result (see Figure 5.2).

Next, categories that are associated with each rated result are recorded in the user profile with a positive or negative rating. When the system has collected enough information about the user in order to provide well-tailored output, relevant hits will appear higher in the list of search results than non-relevant hits. For example, when someone who is interested in 'biology', but not in 'computers' searches on 'ant', search results concerning the ant as an insect should be listed higher than results concerning the Java programming tool called 'Ant'. In short: the most relevant hit for each individual should appear at the top of the list. The Prospector user profile is scrutable (Kay, 2000): users can view and alter it. This enables users to understand the personalization provided by Prospector and to fine-tune the assumptions made by the system in order to optimize tailored output. Adding this feature has been shown to increase feelings of controllability (Kay, 2000).

**Prospector** >> Search

Search     Interests     Profile     Logout 'lex'     About

Sift the web for gems:

art museum vienna

(Search)

Rerank the results as if viewed by somebody interested in  arts  ⬩  (Rerank)

[Show original Google results]

**Museums in Vienna - Vienna Attractions - TripAdvisor**
TripAdvisor Popularity Index: #8 of 228 attractions in **Vienna**. Attraction type: Architectural building; **Art museum**. Traveler Reviews: ...
Recreation / Travel / Guides and Directories /
http://www.tripadvisor.com/Attractions-g190454-Activities-c6-Vienna.html     78%
✖ Unsuitable

**Museum of Art** History / Fine Arts, Vienna
**Museum** of **Art** History / Fine Arts, **Vienna**, tourist attractions, information, pictures, maps.
Recreation / Travel / Guides and Directories /
http://www.planetware.com/vienna/museum-of-art-history-fine-arts-a-w-khm.htm     78%
✖ Unsuitable

Kunsthistorisches Museum - Wikipedia, the free encyclopedia
The Kunsthistorisches **Museum** (English: "**Museum** of **Art** History", also often referred to as the "**Museum** of Fine Arts") in **Vienna**, housed in its festive ...
Computers / Open Source / Open Content / Encyclopedias / Wikipedia /
http://en.wikipedia.org/wiki/Kunsthistorisches_Museum     52%
✖ Unsuitable

Schloss Belvedere Palace & Belvedere Art Museum, Vienna
A Sightseeing Guide with Travel Information for the Schloss Belvedere Palace & **Art Museum**, **Vienna**.
Regional / Europe / Austria / Travel and Tourism / Travel Guides /
http://www.tourmycountry.com/austria/schlossbelvedere.htm     61%
✖ Unsuitable

Kunsthistorisches **Museum** Wien
Das **Museum** und seine Sammlungen werden ausführlich mit qualitativ hochwertigen Abbildungen beschrieben.
World / Deutsch / Kultur / Museen / Bildende Kunst /
http://www.khm.at/     50%
✖ Unsuitable

**Result page**

1 2 3 4 5 6 7 8 9 10 ≫

Copyright © 2006, Institute for Information Processing and Microprocessor Technology (FIM), Johannes Kepler University, Linz Austria.
This system uses data kindly provided by the Open Directory Project.

Help build the largest human-edited directory on the web.
Submit a Site - Open Directory Project - Become an Editor

Figure 5.1. Prospector main screen



◀ Result NOT OK.
Take me back!     ◀ Result OK.
Take me back!     **Google** PROSPECTOR     Result OK.
Stop searching! ▶     Result NOT OK.
Stop searching! ▶

Figure 5.2. Prospector rating frame

### 5.4.2 Experimental procedure

We conducted a user-centered evaluation with think-aloud sessions, interviews and questionnaires in a usability laboratory with 32 undergraduate students of the social sciences. Before starting the evaluation, the participants completed a survey on demographics and computer usage. Each participant was assigned to one of four conditions. Eight participants completed their tasks while thinking-aloud and were subsequently interviewed, while eight completed a questionnaire after thinking out loud. In the other two conditions, participants fulfilled the test tasks without thinking-aloud and were then interviewed (n = 8) or filled out a questionnaire (n = 8). A schematic overview of the participants and their conditions can be found in Figure 5.3. This study design was chosen in order to cope with feasibility restraints: it allowed us to put several methods to the test with a relatively low number of participants. To check for possible consequences of combining methods, we analyzed findings for interaction effects. As will be discussed in the Results section, no interaction effects occurred.
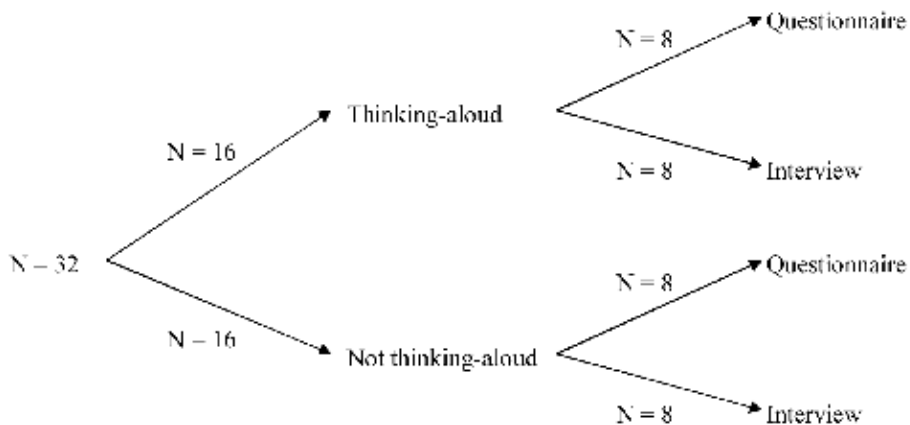


Figure 5.3. Conditions and participants

The first task for participants was to create a user profile in Prospector. To do this, they indicated their interests by means of the system's 'create a login' procedure. Next, they had to use Prospector to search for information on city trips to large European cities. More specifically, they had to search for a youth hostel of their liking and the address of the Museum of Modern Art in four large European cities. They had to write the name of the youth hostel and the address of the Museum of Modern Art down in a booklet. There was a maximum time of ten minutes per search. The Museum of

Modern Art search was what Spool and colleagues (1997) call a simple fact search. The youth hostel search is a comparison and judgment test task. Here, participants have to find relevant information and compare different options. This is the most complicated kind of test task. By choosing two test tasks that differ in their degree of difficulty, comments on interaction with Prospector for different contexts can be elicited. The large freedom the participants had while searching with Prospector increases the reality of the test tasks and consequently, contributes to the diversity and importance of identified usability problems (Cordes, 2001). The participants were encouraged to rate the search results using the rating frame. That way, high quality personalization was made possible. Between the search tasks for the third and fourth city, participants were instructed to look at their user model and alter it to generate a maximum fit between the model and their personal interests.

When a participant was thinking-aloud while completing the tasks, an evaluator sat next to the participant. Before the session with Prospector, the participant was briefed on what thinking-aloud entailed. The evaluators told the participants that they would not answer any question a participant might have. They would remind the participants to think-aloud if necessary. Next, thinking-aloud was practiced by looking up a train schedule on the internet. All think-aloud sessions were audio-recorded. The thinking-aloud data was supplemented with data from observations, when necessary to clarify the audio-recording or when the evaluator identified an action by the respondent that led to lower effective or efficient use of the system.

The questionnaires focused on specific usability issues for personalized systems, appreciation of personalization, and the perceived relevance of search results. These last two issues were assessed by asking the respondent to make a comparison between Prospector and Google (a personalized and a non-personalized search engine). Also, the respondents were asked to give their reasons for using Prospector, or for not using it, where we expected the participants to provide comments on the perceived relevance of search results if he or she formed an opinion about them. All of the questions were open-ended. The interviews posed the same questions as the questionnaire. But since the interviews were semi-structured, the interviewer could ask for clarifications when an answer was unclear. All of the interviews were audio-recorded. One example of a questionnaire and interview item about the specific issue of 'comprehensibility' goes as follows: "Are there certain parts of Prospector that you think are hard to understand? And if so, which ones are they? And why do you think these are hard to understand?"

Figure 5.4 displays the test procedure. The questionnaire and interview items can be found in the Appendix.
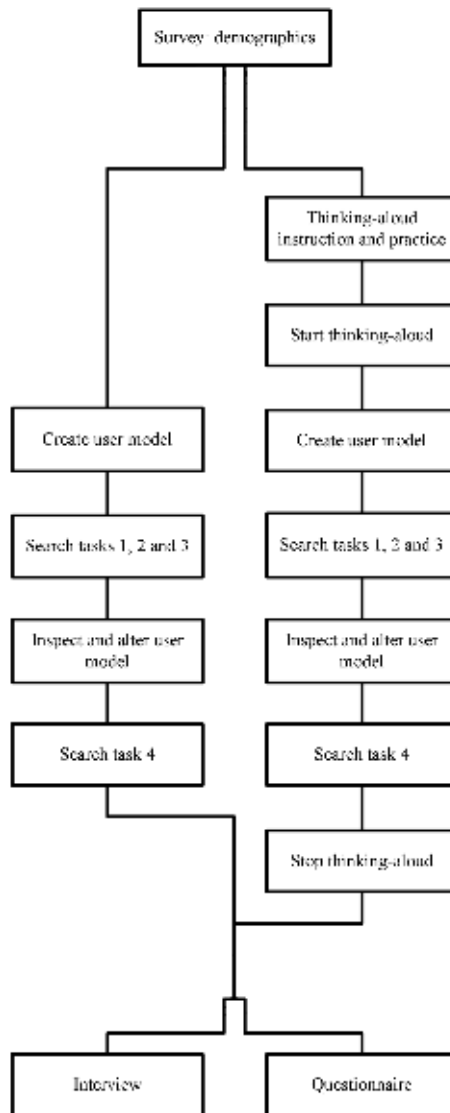
Survey demographics

Thinking-aloud instruction and practice

Start thinking-aloud

Create user model | Create user model

Search tasks 1, 2 and 3 | Search tasks 1, 2 and 3

Inspect and alter user model | Inspect and alter user model

Search task 4 | Search task 4

Stop thinking-aloud

Interview | Questionnaire

Figure 5.4. Study timeline

### 5.4.3 Data analysis

Comments on specific and generic issues were abstracted from the audio recordings or completed questionnaires and transcribed by one of the researchers. When we talk of 'comment', we mean any relevant verbalization of a thought, a thinking-aloud participant's action that had a detrimental effect on effective or efficient use of the system, and relevant feedback pro-

vided during the interview or on the questionnaire. For each comment, the researcher determined the following attributes:

§ Is it a comment on a specific or generic issue?

We classified a comment as specific when it addressed a usability issue specific for personalized systems, appreciation of personalization, or the perceived relevance of search results.

§ If a comment was specific, it was assigned to one of the usability issues for personalization, appreciation of personalization or perceived relevance of search results.

If a comment was not related to any of these categories, it was classified as a generic comment. It was then appointed to one of the problem types, listed by Van der Geest (2004):

− Content & Information problem

− Navigation & Structure problem

− Design & Presentation problem

− Other problem

§ Is the comment positive, negative, neutral, or ambiguous?

We will give one example of a coding to clarify the procedure. During an interview, one participant said: "I don't care about privacy in this particular case: everybody may know about these trivial things I am looking for." This comment was classified as specific, as it addressed privacy, one of the specific usability issues for personalization. Hence, it was also coded as a comment on privacy. Finally, the comment was coded as positive, as the participant stated that he felt Prospector did not infringe on his privacy.

Next, all comments concerning the same problem or popular feature were grouped and named. One example of a problem relating to personalization is "comprehensibility concerning the compilation of the user profile". Here, participants did not understand how the system created their user model. Examples of generic problems are "similarity between Prospector and Google" or "understanding of keywords" in the user model.

In order to determine the value of different evaluation methods, it is also important to make a distinction between the severity of the different problems that have been identified (De Jong & Schellens, 2000; Hartson, Andre, & Williges, 2001; Hornbæk, 2010). Therefore, we concentrated on the negative comments, and in line with (Høegh & Jensen, 2008; Hornbæk & Frøkjær, 2005; Kjeldskov & Stage, 2004), we classified problems as critical, serious or minor. The following definitions are derived from Duh, Tan, & Chen (2006). A critical problem prevented participants from completing tasks and/or recurred across all participants. A serious problem severely

increased the task completion time and/or recurred frequently across participants. However, a serious problem did not prevent a participant from completing the task eventually. A minor problem increased task completion time slightly and/or recurred infrequently across the evaluation participants. Finally, a minor problem did not prevent the evaluation participants from completing a test task easily.

### 5.4.4 Pitfalls of evaluating personalization

Whenever a personalized system is evaluated, several considerations have to be taken into account to ensure the value and validity of the evaluation. In this section we will list the most important ones and how we have dealt with them.

It might be difficult for participants to give their opinion on personalization (Weibelzahl, 2005). They might only notice the effect of personalization when the output does not match their characteristics, preferences or context. When personalized output *does* provide a clear match, personalization might go unnoticed. This complication makes it clear that, during an evaluation, the perceived quality of personalization should not be asked about directly (e.g., "Do you like the personalization being done?"). It should be posed in terms of the variable it is supposed to serve. For example, when one is evaluating an e-learning system that provides personalized instructional texts one should not ask "Do you like the personalization of the instructional text you have just read?" Rather, the question should be: "Did this text help you to achieve your learning goal?" This last example refers to the perceived quality of personalization in relation to the goal of personalization. We defined the success of personalization as positive 'perceived relevance', a measure of search engine success suggested by Nahl (1998). Therefore, 'perceived relevance of search results' was coded as a result of personalization in our data analysis.

Many personalized systems are faced with the so-called *cold start problem*. The degree of personalization presented to the user increases during user-system interaction. Many personalized systems start with no personalization at all and 'learn' about the user during interaction. This knowledge is then used to tailor output. Therefore, adaptive systems require an investment by the user in order to create personalized output (Höök, 1997). For the evaluation of personalized systems, this means that it needs to be provided with user information before a session (if full personalization is needed from the beginning), or the session must offer the possibility for the system to create a complete and valid user model. The latter will require a certain amount of interaction which lengthens session times. In this study, the *cold*

*start problem* was accounted for by using two test tasks in a single domain and repeating these two tasks four times. This way, the system could be personalized for one search domain in a relatively short time.

The evaluation of a personalized system should be conducted (even more so than in the case of non-personalized systems) with a heterogeneous group of participants. A personalized system should provide meaningful, personalized output to every user in every context and the quality of this output can only be evaluated fully if many different users with different contexts are represented (Weibelzahl, 2005). Therefore, the applied methods should account for the participants' context (Akoumianakis, Grammenos, & Stephanidis, 2001). Our group of participants (students) was homogeneous not heterogeneous, which limits the diversity of results. However, in this study our goal is to identify differences among different user-centered evaluation methods, and not to generate an exhaustive list of usability issues for Prospector. By using a homogeneous group of participants, our results can only be attributed to the different methods and not to the different user populations that were present in our study. That is why we decided to use a homogeneous population of students as participants.

## 5.5 Results

### 5.5.1 Participants

In total, 32 Social Science undergraduates participated in the evaluation. Twenty-four of them were female and eight were male. They had an average age of 19.9 years (SD = 2.0 years). The participants used a computer and the internet on a daily basis and were familiar with some personalized systems, such as:

§ Amazon's book recommendations (11 participants);
§ Bol.com recommendations (9 participants);
§ My IB-group (the Dutch personalized website on student loans; 19 participants): and,
§ *i*Google (11 participants).

On average, the thinking-aloud participants took 48 minutes and 40 seconds to complete the test tasks. The other participants took on average 39 minutes and 4 seconds.

### 5.5.2 Quality of measurement

First we have to determine the reliability of the coding of comments: inter-coder agreement. Therefore, an external usability expert re-coded a subset of the data independently; an approach suggested by Gray and Salzman (1998).

According to intercoder-reliability guidelines, 10% of the comments were re-coded with a minimum of 50 comments. Therefore, a total of 56 user comments on specific issues and 50 user comments on generic issues were coded again. Next, Cohen's Kappa was calculated for each variable. The average Kappa score was .73 which, according to Byrt (1996), stands for 'good agreement'.

Table 5.1 Number of comments, and their valence, elicited by questionnaires and interviews

|  | Comment type | TA* first | | no TA first | |
|---|---|---|---|---|---|
|  |  | mean | SD** | mean | SD |
| Questionnaire | positive | 4.00 | 2.33 | 3.88 | 2.90 |
|  | negative | 4.62 | 2.88 | 5.12 | 3.56 |
|  | neutral | 1.38 | .74 | 1.75 | 1.75 |
| Interview | positive | 5.38 | 2.39 | 7.50 | 2.14 |
|  | negative | 562 | 3.02 | 3.62 | 2.77 |
|  | neutral | 3.75 | 1.98 | 2.50 | 1.51 |

* TA = Thinking-aloud
** SD = Standard deviation

Table 5.2 Number of comments, related to personalization, elicited by questionnaires and interviews

|  | Comment type | TA* first | | no TA first | |
|---|---|---|---|---|---|
|  |  | mean | SD** | mean | SD |
| Questionnaire | specific usability issues | 8.75 | 1.83 | 9.12 | 1.96 |
|  | appreciation of personalization | 1.00 | .76 | 1.00 | .93 |
|  | perceived relevance of search results | .25 | .46 | .62 | 1.41 |
| Interview | specific usability issues | 12.00 | 3.78 | 11.38 | 3.54 |
|  | appreciation of personalization | 1.38 | .74 | 1.62 | .84 |
|  | perceived relevance of search results | 1.38 | 1.06 | .62 | .92 |

* TA = Thinking-aloud
** SD = Standard deviation

During the evaluation, there were two groups: participants who did think out loud and participants who did not. One can ponder whether this influenced the number of answers they gave during on the questionnaires or during interviews. Did the participants who thought out loud first give more answers during these sessions than the participants who did not think aloud first, or vice versa? The average number of each type of comment gathered by the interview or questionnaire with or without being preceded by thinking-aloud can be found in Table 5.1 (specified for valence of the comments)

and Table 5.2 (specified for comments related to personalization). We combined the comments given on specific usability issues into one category, as the analysis of each issue separately would not have resulted in meaningful results: The number of comments on each separate usability issue elicited by the questionnaire or the interview appeared to be too low.

We tested whether there was a significant difference between the number of comments of each type gathered by questionnaires that were preceded by thinking-aloud or not, by means of t-tests. These tests showed that the number of comments (for each category of valence, or category related to personalization) collected by the questionnaire preceded by thinking-aloud or not, did not differ significantly. The same result was found for the interview: The number of comments for each category collected by the interview, either preceded by thinking-aloud or not, were the same.

These results show that our data set provided us with a good basis to compare the yield of the three user-centered evaluation methods for evaluating personalization.

### 5.5.3 Comments on personalization

The questionnaires, interviews and thinking-aloud sessions elicited 555 comments on personalization altogether. Questionnaires accounted for 227 of them, interviews for 166 and thinking-aloud sessions for 162. Of these 555 comments, 179 were positive, 91 were neutral or ambiguous, and 285 were negative.

## 5.5.3.1 Specific usability issues

If applicable, we determined the specific usability issue to which a comment could be attributed. Table 5.3 shows how many times a comment on each specific issue was made in each questionnaire, interview, or thinking-aloud session. One example of a comment on the breadth of experience goes:

Interviewee: "I know I missed information by using Prospector. I know of sites with lists of youth hostels which come in handy. With Google I get them all the time, but not with Prospector."

One participant commented on the comprehensibility of Prospector on the questionnaire:

"The more I use Prospector, the more information concerning my interests is stored. This way, the program can use my interests to adjust search results to me as an individual."

Table 5.3. Number of comments on specific issues per session

| | Questionnaire | | Interview | | Thinking-aloud | |
|---|---|---|---|---|---|---|
| | mean | SD* | mean | SD | mean | SD |
| Predictability | 1.00 | .37 | 1.19 | .54 | .69 | .87 |
| Comprehensibility | 2.19 | .75 | 2.19 | 1.05 | 1.50 | 1.32 |
| Controllability | $.75^{T}$ | .45 | $1.13^{Q/T}$ | .50 | .06 | .25 |
| Unobtrusiveness | $1.06^{T}$ | .68 | $1.56^{T}$ | 1.03 | .06 | .25 |
| Privacy | $1.50^{T}$ | .63 | $1.63^{T}$ | 1.03 | .06 | .25 |
| Breadth of experience | 1.44 | .89 | $2.19^{T}$ | 1.17 | 1.00 | .97 |
| System competence | 1.00 | .63 | $1.81^{T}$ | 1.38 | .89 | .50 |

\* SD = Standard deviation

Note: Letters in superscript behind mean indicate that this mean is significantly higher than the mean of the method where Q = Questionnaire; I = Interview; T = Thinking-aloud

ANOVA analyses were conducted to ascertain whether the number of comments yielded by each method differed. The ANOVA analyses uncovered significant differences for the number of comments on:

§ controllability ($F(2,45) = 27.20$, $p < .01$, $\omega = .79$);

§ unobtrusiveness ($F(2,45) = 17.64$, $p < .01$, $\omega = .71$);

§ privacy ($F(2,45) = 23.93$, $p < .01$, $\omega = .77$);

§ breadth of experience ($F(2,45) = 5.60$, $p < .01$, $\omega = .47$); and,

§ system competence ($F(2,45) = 6.80$, $p < .01$, $\omega = .52$).

No significant differences were found in the case of predictability ($F(2,45) = 2.57$, $p > .05$, $\omega = .25$) and comprehensibility ($F(2,45) = 2.23$, $p > .05$, $\omega = .20$). Table 5.3 shows that people do not comment on the topic of predictability in all three conditions. Comprehensibility is the only topic on which thinking-aloud elicited some comments, thereby preventing a significant difference with interviews and questionnaires from occurring on this one issue. For the five issues with significant differences we conducted post hoc analyses by means of Bonferroni tests at a 5% significance level. The results can be found in Table 5.3. Interviewing resulted in more comments on these five issues than thinking-aloud. In the case of controllability, the interview elicited more comments than the questionnaire. The questionnaire provided more comments on controllability, unobtrusiveness and privacy than thinking-aloud. Thinking-aloud supplied only a marginal number of comments on the specific usability issues for personalization.

### 5.5.3.2 Appreciation of personalization and the perceived relevance of search results

Table 5.4 shows how many comments on the appreciation of personalization and perceived relevance of search results each method yielded. One example of a comment on the appreciation of personalization is:

Interviewee: "I don't like the personalization of search results. You don't always want to search the same thing and with the same line of approach. I think it's a useless feature."

One thought expressed about the perceived relevance of search results went as follows:

Thinking-aloud: "It strikes me that there are a lot of Irish and New York museums [in my search results], while I am looking for museums in Hamburg."

Table 5.4. Number of comments on appreciation of personalization and perceived relevance of search results per session

|  | Questionnaire | | Interview | | Thinking-aloud | |
|---|---|---|---|---|---|---|
|  | mean | SD* | mean | SD | mean | SD |
| Appreciation of personalization | $1.00^T$ | .82 | $1.50^T$ | .73 | .13 | .34 |
| Perceived relevance of search results | .44 | 1.03 | 1.00 | 1.03 | $6.13^{Q/I}$ | 2.36 |

* SD = Standard deviation

Note: Letters in superscript behind mean indicate that this mean is significantly higher than the mean of the method where Q = Questionnaire; I = Interview; T = Thinking-aloud

ANOVA analyses uncovered differences among the number of comments on both appreciation of personalization ($F(2,45) = 17.66$, $p < .01$, $\omega = .71$) and the perceived relevance of search results ($F(2,45) = 61.13$, $p < .01$, $\omega = .89$). Again, we conducted post hoc analyses by means of Bonferroni tests at a 5% significance level. The results can be found in Table 5.4. When it comes to collecting comments on the appreciation of personalization, both the questionnaire and the interview were more useful than thinking-aloud. In the case of perceived relevance of search results, the opposite picture emerged. In that case, thinking-aloud turned out to be far more useful than the questionnaire and the interview.

The results we found concerning comments on the usability issues for personalized systems, appreciation of personalization, and perceived relevance of search results do no support hypothesis 1 (Thinking-aloud, ques-

tionnaires and interviews yield the same number of comments from participants on specific usability issues and appreciation of personalization), but do support hypothesis 2 (Thinking-aloud elicits more comments from participants on the perceived relevance of search results than the questionnaires and interviews).

### 5.5.3.3 Positive, neutral, and negative comments

We coded comments for their valence: positive, negative, or neutral or ambiguous. One example of a comment that was coded as 'ambiguous' was "I don't think Prospector is an improvement over Google, but it does add something." Table 5.5 shows the differences in the number of positive, neutral or ambiguous, or negative comments that were elicited.

Table 5.5. Number of differently valued comments on personalization per session

|  | Questionnaire | | Interview | | Thinking-aloud | |
|---|---|---|---|---|---|---|
|  | mean | SD* | mean | SD | mean | SD |
| Positive | $3.94^{T}$ | 2.54 | $6.44^{Q/T}$ | 2.45 | .81 | 1.05 |
| Neutral | 1.56 | 1.32 | $3.13^{Q/T}$ | 1.82 | 1.00 | 1.21 |
| Negative | 4.88 | 3.14 | 4.63 | 2.99 | $8.31^{Q/I}$ | 2.75 |

* SD = Standard deviation

Note: Letters in superscript behind mean indicate that this mean is significantly higher than the mean of the method where Q = Questionnaire; I = Interview; T = Thinking-aloud

We performed ANOVA analyses to see whether the number of the differently valued comments each method yielded differed. There appeared to be significant differences in all cases, i.e., positive issues ($F(2,45) = 28.13$, $p < .01$, $\omega = .79$), neutral issues ($F(2,45) = 8.94$, $p < .01$, $\omega = .58$) and negative issues ($F(2,45) = 7.74$, $p < .01$, $\omega = .54$). Next, we conducted Bonferroni tests with a 5% significance level to find out which groups differed. The results of these post hoc analyses can be found in Table 5.5. Both the questionnaire and the interview elicited more positive comments than thinking-aloud. The interview elicited more positive comments than the questionnaire. Furthermore, the interviews resulted in more neutral comments than the questionnaire and thinking-aloud. Thinking-aloud ultimately supplied more negative comments than the questionnaire and the interview. A large part of these comments consisted of remarks on negative perceived relevance of search results and the participants' rationale for this negative perception. On average, 5.38 of such comments were made per thinking-aloud session.

These results support our third hypothesis: Thinking-aloud elicits more negative comments on personalization than questionnaires and interviews.

## 5.5.3.4 Problem severity

After analyzing the collected comments with a quantitative approach, we looked at their actual content. By grouping and naming negative comments on the same problem, we were able to see whether the different methods detected the same or different problematic issues. The Venn diagram in Figure 5.5 displays each method's contribution to the total set of identified problems. Here, we can see that each method contributes a unique set. Furthermore, a considerable number of issues were identified by two of the methods. The number of issues identified by all three methods was relatively small.
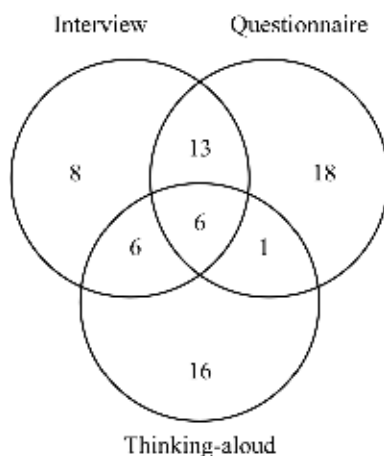


Figure 5.5. Problematic issues related to personalization uncovered by each method

We identified 2 critical, 13 serious and 53 minor problems. The quality of coding of problem severity was again assessed using an external usability expert who re-coded the problems, according to the guidelines described before. A comparison of the original and re-coded dataset resulted in a Cohen's Kappa of .76, which stands for 'good agreement'.

One critical problem was identified by thinking-aloud only. The other critical problem was mentioned by participants who were thinking-aloud or who were interviewed. It is worth mentioning that this problem was brought forth three times during an interview and 22 times during a thinking-aloud session. Figures 5.6 and 5.7 show how each method has contributed to the set of serious and minor problems that were identified. In the case of serious problems, thinking-aloud uncovered two issues that were not identified by other methods. The other problems were identified by two, or all three

methods. Finally, each method produced a unique set of problems. Regarding minor issues, the set identified by the questionnaires was the largest, followed by the set that resulted from the thinking-aloud sessions. The interviews uncovered a relatively small set of eight minor problems.
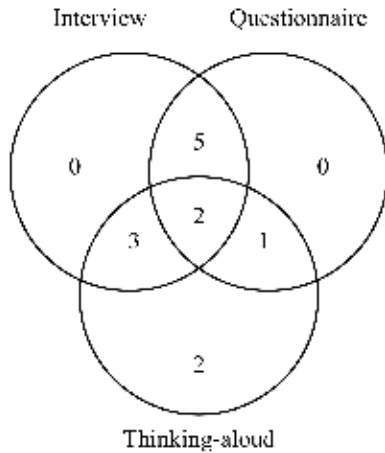


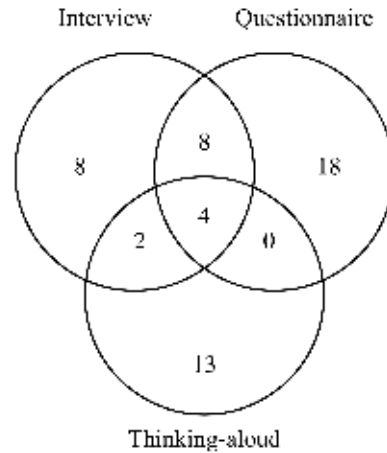Figure 5.6. Serious problems relating to personalization uncovered by each method



Figure 5.7. Minor problems relating to personalization uncovered by each method

Our fourth hypothesis (The problems related to personalization identified by thinking-aloud on the one hand, and questionnaires and interviews on the other, do not overlap) is partly supported by these results. The Venn diagrams show that there is a certain overlap between the problems the methods identified. However, only by means of thinking-aloud could both critical and two serious problems be uncovered. Interviews or questionnaires did indeed elicit a set of serious problems that thinking-aloud did not, but the application of both methods was not necessary. The use of only interviews or questionnaires in combination with thinking-aloud would have resulted in the same list of serious problems. Finally, minor problems were relatively rarely elicited by two or all three methods. In this case, each method has a unique yield.

### 5.5.4 Comments on generic issues

Besides the comments on personalization, the thinking-aloud sessions resulted in 159 comments on generic usability issues. Of these comments, 108 were negative, 35 were neutral or ambiguous, and 16 were positive.

Each comment on a generic usability issue was placed in one of the comment categories, listed under Van der Geest (2004): content & information, navigation & structure, design & implementation, and other comments. The number of comments in each category, which differed significantly from each other ($\chi^2(3, N = 159) = 56.17$, $p = .00$), are displayed in Table 5.6. It shows that most comments were made on the content and information provided by Prospector (76 in total). About half of that number of comments was directed at the navigation and structure (39), and the design and presentation (34) of the system.

Table 5.6 Topics of generic issues commented on during thinking-aloud

|  | Content & information | Navigation & structure | Design & presentation | Other | Total |
|---|---|---|---|---|---|
| Positive | 7 | 2 | 2 | 5 | 16 |
| Neutral | 11 | 3 | 19 | 2 | 35 |
| Negative | 58 | 34 | 13 | 3 | 108 |
| Total | 76 | 39 | 34 | 10 | 159 |

In a similar way as we did for the specific issues, we looked at the content of the comments on generic comments and grouped them. Then, we designated a severity rating to each problem: critical, serious or minor. These severity ratings were in line with the definitions given beforehand. Again we calculated Cohen's Kappa: .75. In total, there were 36 problems, of which 30 were minor, 5 were serious, and 1 was critical. The critical problem was brought forward by 12 participants who did not understand the interest categories displayed in their user model. As a result, they were unable to alter it correctly. One participant, for example, when confronted with the category 'France' in her user model said:

"'France'? Does that mean only sites in French? Or something different?"

Although the understanding of labels used on an internet page can be considered to be a generic usability issue, it can have major detrimental effects on the quality of personalization. If users do not correctly indicate their interest in special interest categories or keywords, future personalized output may not be in line with their specific, characteristics, preferences and context.

# 5.6 Conclusions and discussion

## 5.6.1 Uncovering specific issues

Thinking-aloud has an important function during formative evaluations, namely as the supplier of a large number of comments on the perceived quality of personalized output. For a personalized search engine, perceived relevance is the most important variable in an evaluation. It determines for a large part how useful such a system will be, and the extent to which it will be used (Tsakonas & Papatheodorou, 2006). These comments appeared to be elicited best by the method that collects users' thoughts 'on the fly.' Therefore, thinking-aloud should be considered a crucial part of the formative evaluation of a personalized system. Not only does it show whether personalized system output is perceived as appropriate or not, it also tells one *why*. Such information is of great value for system redesign. The importance of thinking-aloud is supported by the results on the different contributions of each method to the collection of critical and serious problems concerning personalization. In a formative stage of the design process (a stage in which an evaluator wants to identify issues that need to be improved), one cannot do without thinking-aloud, as it is the only method that unearthens all the critical problems as well as several serious problems.

When we focus on the specific usability issues for personalization and the appreciation of personalization, it appears that the questionnaire and the interview are more suitable for generating comments than thinking-aloud. Issues like privacy, predictability, etc., are of a more general nature. Participants only seem to be able to comment on them when they are explicitly asked to consider these issues.

There are several differences between the potential of the questionnaire and the interview with regard to providing the evaluator with an impression of how specific issues are experienced. First, the interview yields more comments on the issue of controllability than the questionnaire while both methods elicit more comments on specific usability issues than thinking-aloud (for the specific usability issues of controllability, unobtrusiveness and privacy; the interview also elicits more comments on the breadth of experience and system competence). So when one wants to receive the largest number of comments on usability issues for personalization and the appreciation of personalization, one should select the questionnaire before thinking-aloud, and the interview before the questionnaire. The higher number of comments collected by the interview may be the result of the fact that the number of comments collected per question in the questionnaire tends to be just one.

This means that participants do not engage in elaborate answers when completing a questionnaire. However, the interview collected significantly more positively comments, which suggest that the interview may be positively biased. It might be that participants want to save face (for example, when they were asked whether they understood how the system works), give socially desirable answers, or please the experimenter. One final point is that the questionnaire, even after careful pretesting, can pose a question which the participant does not understand. In this case, the evaluator will not receive the kind of answer that was hoped for. This was the case in our questionnaire where the topic of controllability received an average amount of .75 comments, although a question explicitly asking for a comment on it was included. When a participant does not understand a question while being interviewed, this matter can simply be resolved by the interviewer.

### 5.6.2 Uncovering generic issues

The results of this study underline the suitability of thinking-aloud for identifying unsatisfactory features or system output, as previously stated by Benbunan-Fich (2001). Thinking-aloud elicited most comments on content and information, followed by a smaller number of comments on navigation & structure, and design & presentation. However, the distribution of the comments over the comment categories may have been influenced by the system under evaluation. Two specific problems accounted for almost half of the comments on content & information, and therefore may have, unjustly, painted a picture that thinking-aloud yields more comments on this specific topic. However, thinking-aloud provides the evaluator with insights on the (critical) generic usability issues of a personalized system. These issues may well have detrimental effects on the quality of personalized output. So, using thinking-aloud to identify generic usability problems in a personalized system is crucial for improving the system.

### 5.6.3 Limitations of this study

The specific system under investigation, Prospector, may have influenced the distribution of user comments. So what is the generalizability of the findings of this study to evaluation in general? In the case of generic usability issues, which we have so far reported on in a general way, the results are indeed heavily influenced by the issues present in the system under investigation, as in any usability study. For the issues related to personalization, we think that the conclusions hold for formative evaluations of systems that apply a similar form of personalization as Prospector: link sorting. The system may have influenced the number of comments per issue on personaliza-

tion, but this does not affect the general trend over multiple methods. Examples of such similar forms of personalization include altering text fragments (as in personalized recommendations), or link annotation (where personally important links on a website stand out by using divergent colors or fonts). These techniques have a similar approach as they create a personal lay-out of text. They also have a similar goal as they aim to guide the user to personally relevant information. Whether or not the results hold for a formative evaluation of personalized systems in general, is difficult to say. Other forms of personalization may have a different approach and goal. Link hiding, for example, is a form of personalization that needs to be evaluated differently. People might not notice that something is being personalized, as they cannot see the personalization being done. As a result, thinking-aloud sessions may be useless here. In order to find out which evaluation methods are best suited to evaluate other forms of personalization, future studies using systems that apply a different personalization technique are necessary.

The interview and questionnaire items used in this study were just a selection of all the possible items one can create. The items one uses influence the kind of comments elicited from participants as they guide the participants' line of thought. Therefore, it may be possible that other items may have elicited other kinds of comments. We could have tested this by using multiple questionnaires or interview schemes. However, feasibility constraints prevented us from doing so as it would have necessitated a larger group of participants. We are of the opinion that the items we formulated were well-suited for eliciting the desired comments. At the same time, we realize that using interview and questionnaire items that have been optimized after several rounds of testing may have led to different results. However, such items are currently not available. In the future, it would be worth replicating this study with the same items to confirm the results we found, or to use items that are the result of iterative questionnaire or interview design to see whether other questionnaires or interviews might yield different results.

### 5.6.4 The integrated, formative evaluation of personalization

The results of this study will encourage evaluators to apply both thinking-aloud and questionnaires in the formative evaluation stage of a personalized system. This dual approach elicits the most important problems and does so in a valid way. If one is primarily interested in the effectiveness and efficiency of the system (as is the case in the summative evaluation stage), other user-centered evaluation methods may have a bigger yield. Thinking-aloud, for example, may be less suitable in this instance, as thinking-aloud may

require mental effort from participants which can lengthen the time they need to complete test tasks (Hertzum, Hansen, & Andersen, 2009; Holzinger, 2005).

A formative evaluation of a personalized system will never focus solely on specific or generic usability issues. The two are inseparable. For example, if a user does not improve his/her user model correctly because he or she cannot operate the visualization technique applied in the interface, one might be tempted to say that there is only a generic usability problem. However, as a result, the system may store an incorrect assumption about the user and utilize it to generate personalized output. This way, the problem influences the quality of the personalized output and can become a specific issue. It would, therefore, be too limiting to evaluate a personalized system exclusively from a personalization or a generic usability perspective.

We would like to stress that the array of user-centered evaluation methods is much larger than merely questionnaires, interviews and thinking-aloud sessions. We assume that other methods (e.g., expert reviews) can also contribute positively to the evaluation of a personalized system. It would be worthwhile comparing more methods systematically in order to understand and fine-tune the full range of possibilities that evaluators of personalized systems have at their disposal.

*In chapter 2 to 5 I have discussed studies related to four phases in the user-centered design process. These studies have resulted in concrete design guidelines for personalization or insights in the value of different methods for requirements engineering or formative evaluations, when applied to personalization.*

*In the final chapter of this thesis, I will summarize the main findings of these studies and discuss how the user-centered design approach can be of added value for personalized system design in the future.*

# Chapter 6

## Reflection

## 6.1 Main findings of this thesis

In chapter 1, I introduced the concept of personalization and showed how tailored electronic communication is the product of centuries of evolution. Personalization involves gearing communication towards an individual's characteristics, preferences and context. User-Centered Design (UCD) was proposed as a means to achieve a good fit between personalized communication and the individual user. This means that design of personalization should include an initial focus on users and their tasks, studies should be conducted that focus on actual user behavior and perceptions, and finally, an iterative design approach should be applied. In this way, problematic issues related to specific, personalized usability issues, such as privacy or a need for control, can be prevented.

Chapter 2 addressed an early stage in the UCD process of personalization to determine the role of trust in the organization providing personalization, trust in the technology, and perceived controllability in relation to the intention of potential users to use online content personalization. Using an online questionnaire, 1,141 participants were demonstrated four common approaches to online content personalization and a non-personalized baseline condition with respect to a fictive municipality. We assessed participant perceptions of the aforementioned factors and determined their influence on the intention to use the different approaches to online content personalization. Trust in the organization appeared to play no role in the decision to use online content personalization. Trust in the technology had a moderate effect on the intention to use, while perceived controllability was overall the most important antecedent. When designing online content personalization, it is therefore most important to provide users with the option to control personalization. Next, users should be assured that they are interacting with an organization in a secure electronic environment.

The requirements engineering phase was focus of chapter 3. In that chapter, we proposed a user-centered approach to requirements engineering for personalized e-Government services and demonstrated its value by means of a case study. The approach utilized interviews and formulated requirements by focusing on concrete and measurable criteria, low-fidelity prototyping, and evaluating by means of a citizen walkthrough. The case study reaffirmed the importance of applying an iterative approach to design, as the translation of user input into system design may not align with the original characteristics, preferences and contexts of the user. Furthermore, using a citizen walkthrough, the proposed approach succeeded in making

personalization understandable to participants, which is an important objective for evaluating personalization. Finally, the case study demonstrated that a multidisciplinary design team is a crucial aspect of creating personalized e-Government services.

In chapter 4, we reviewed literature that focused on user-centered evaluation of personalization (i.e., evaluations that include an assessment of subjective criteria or the identification of usability problems). The findings indicate that current user-centered evaluations, as reported in the scientific literature, are not well-aligned with the principles of UCD. Questionnaires appeared to be exceedingly popular, while methods that have been found to identify usability problems well, such as thinking-aloud techniques, are only used sparingly. Specific usability issues for personalization are only rarely a topic of investigation. In the last few years, however, an increasing number of publications have reported on evaluations that focus on acceptance, iterative design or system trust. This trend suggests that personalization researchers are becoming aware of the added value of user-centered evaluations and are starting to make it part of their common research practice.

Chapter 5 reported a comparison of the usefulness of three methods (i.e., interviews, questionnaires with open-ended questions and concurrent thinking-aloud techniques) for identifying usability issues in personalized systems. Thinking-aloud was the only method that uncovered all critical and serious problems related to personalization as well as usability problems not related to personalization. Furthermore, it was also the method that best elicited participant feedback on the perceived quality of personalized output. Comments on the specific usability issues for personalization were elicited best by the questionnaire. Therefore, when evaluating a personalized system in order to obtain input for redesign purposes, we recommend a combination of thinking-aloud techniques and questionnaires with open-ended questions that address specific usability issues in personalization.

## 6.2 User-centered design and layered evaluation

The focus of this thesis has been on UCD. Of course, this is not the only approach available with respect to designing personalization. One technical approach that is receiving increasing attention in from personalization researchers is so-called 'layered evaluation'.

The premise of this approach is that design or evaluation activities related to personalization should not be oriented toward the personalization process as a whole but should break it down into several steps so as to make it possible to pinpoint and solve problems (Brusilovsky, Karagiannidis, &

Sampson, 2001; Paramythis & Weibelzahl, 2005). Each step can then be designed or improved after assessing its validity or reliability separately. Such design and evaluation activities should have the goal of minimizing errors while interpreting information about the user, or deciding upon suitable personalized output. Paramythis, Weibelzahl and Masthoff (2010) have divided personalization into the following steps.

1. Collection of input data
2. Interpretation of collected data
3. Modeling the current state of the "world"
4. Personalization decision
5. Application of personalization

The first step involves *the collection of input data*. Input data comes mainly from user behavior vis-à-vis the system or from explicit user input. Design and evaluation should focus on the correct collection of this data, and as such, this is mostly a technological endeavor (Paramythis, Weibelzahl, & Masthoff, 2010). This step addresses questions such as are keystrokes recorded correctly, or, is the tracking of a user's eye movements precise enough?

UCD can contribute to this step in several ways. First, users should be *willing* to provide the input data. User-centered designers can identify the factors that contribute to such willingness and translate this knowledge into personalized system design. Second, when user data are explicitly collected (e.g., via an e-form), users should be *able* to provide this data correctly. This means that the associated interface and interaction design should be usable. As the design and evaluation of usable technology are traditionally areas of expertise for user-centered designers, their contribution can be of great value here.

The second step, *interpretation of the collected data*, focuses on making sense of the collected data. Most often, this entails making inferences about the user based on the collected data (Paramythis, Weibelzahl, & Masthoff, 2010). For instance, the printing of a web site with the latest news on the Tour de France can be taken as an indication of a user's interest in cycling. The layered evaluation approach encourages conducting design activities that focus on uncovering similar valid assumptions for a given system as well as making evaluations that determine and improve the quality of these assumptions in practice.

The third step involves *modeling the current state of the "world"*. In this step, the interpreted data is stored in the user model. In some cases, this also involves another round of interpretation by the system of the interpreted data (Paramythis, Weibelzahl, & Masthoff, 2010). As in the previous step,

design and evaluation activities should focus on validity as the system performs another round of interpretation.

In the second and third step, user-centered designers can be of assistance when determining *which* data should be used and how this data should be *interpreted*. Ideally, the selection of data and interpretation rules are based on user studies and are evaluated at the earliest possible stage. User-centered designers can help in the design of user studies, user evaluations and the analysis and interpretation of results.

In the fourth step, a system must *make decisions regarding personalization*. It must be decided whether personalization is necessary, whether the selected personalization strategy is appropriate, and whether the applied personalization strategy is acceptable for the user (Paramythis, Weibelzahl, & Masthoff, 2010). This means the user must perceive personalization as useful and easy to use, and it must be in accordance with his or her preferences regarding issues such as controllability, privacy, and so on. Design and evaluation activities should ensure that the selected personalization strategy has a good fit with the user's subjective feelings.

User-centered designers can investigate whether personalization in a certain context is *effective and efficient from a user's point of view*. Furthermore, studies on *user acceptance* can inform the design of personalization so that it takes into account user preferences on important issues in personalization such as privacy.

In the fifth and final step, *application of personalization*, the user is confronted with personalization. Here, it is important that design and evaluation activities ensure that the presentation of tailored output and the design of related interactions are free from usability issues (Paramythis, Weibelzahl, & Masthoff, 2010). Both traditional usability issues and specific usability issues for personalization play an important role here. User-centered designers can play an important role in the design and evaluation of the presentation of personalized output and the design of related interactions.

When all these separate steps for a given personalized system have been designed and evaluated successfully, a personalized system should function well. At that point, it is ready for launch and summative evaluation.

The empirical studies in this thesis have enlarged the toolkit of designers and evaluators using the layered approach in several ways. The findings reported in chapters 2, 3 and 4 are relevant for *deciding on personalization* and *applying personalization*. Chapter 2 pointed out which forms of online content personalization are most appreciated by users and which acceptance-related factors are important in that context. Chapter 3 and 4 discussed effective methods for the engineering requirements in personalized e-

Services as well as suitable methods for user-centered evaluations of personalization. These methods take into account the necessity, appropriateness and subjective acceptance of personalization (Chapter 3) as well as its usability, performance and appropriateness with respect to users (Chapter 4). Chapter 5, finally, reports on the suitability of the three methods for identifying (specific) usability issues and the subjective acceptance of personalization (i.e., *deciding on personalization* and *applying personalization)*. This chapter also revealed how evaluators can determine problems that originate in the *interpretation of the collected data* and/or the *modeling of the current state of the "world"*, which might lead to erroneously personalized output. We have thus identified a method that is suitable for evaluating a transparent user model.

User-centered design and layered evaluation may have a different origin (that is, technical versus user-focused orientations), but both approaches can be of great value. Ultimately, the two approaches should be integrated into a single design process so as to maximize the chances of success for a personalized system. By applying an integrated approach, personalized features can be created that cater to the wishes, need and unique contexts of users; such features should also be of high technological quality, and an integrated approach is best able to make correct assumptions about users and, consequently, make valid decisions for users on the basis of these assumptions.

## 6.3 Personalization: The Holy Grail?

One of the steps in the layered evaluation approach deals with *deciding upon personalization*. This step raises a question that needs to be answered in a very early stage of system design. Is it necessary to implement personalized features in a given system? Or perhaps the use of this technology does not contribute to, or even has a detrimental effect on, the interaction between the user and the system?

It has been argued, especially in the e-government literature, that personalization is a more mature form of online service delivery and by definition is better than non-personalized service delivery (Andersen & Henriksen, 2006). Likewise, adaptivity has been seen as a more advanced level of personalization as compared to adaptability, while both approaches can be seen as more advanced than systems that do not apply any personalization at all (Paramythis, 2009). Finally, according to Plato's view on tailoring communication, as posited in *Phaedrus* (see Chapter 1), one should strive for personalization in communication by profiling an individual and then tailoring communication on the basis of this profile.

Several findings in the studies reported in this thesis do not support the view that (a more advanced form of) personalization is by definition better for users. First, as described in chapter 2, the intention to use online content personalization is not greater when a more advanced form of personalization is available. In contrast, a non-personalized or adaptable form of content provision appeared to be more promising as such content provisions allow the user to have a higher degree of control over the communication process. Second, chapter 3 noted that a personalized approach to finding home help for social support clients was not seen by all prospective clients as a useful substitute for the traditional, offline approach to seeking a help: they valued face-to-face contact too much. Third, chapter 5 revealed the value of different evaluation methods for the formative evaluation of personalization by means of a case study on personalized search. Participant comments as well as results from a longitudinal user study on a subsequent version of a personalized Internet search engine (see Van Velsen, König, & Paramythis, 2009) indicated that personalization in this context does not always add value. More specifically, when searching for information to solve a question with a simple, clear-cut answer (e.g., what is the capital of Uruguay?), personalization did not seem to have an added value. When searches were of a more explorative nature (e.g., searching for information on Asian literature) or were aimed to answer a question for which there is more than one correct answer (e.g., finding a suitable hotel in Barcelona), the use of personalization appeared to be more promising.

Rather than claiming that personalization is always better than no personalization and that adaptivity is better than adaptability, it appears that *a certain form of* personalization *can* be better than no personalization, *depending on the task that is to be performed*. The use of personalization should not automatically be assumed to be value-adding. Its use in relation to the task at hand should be duly considered before creating a system, and designer assumptions about the usefulness of personalization for a given task should be evaluated as early as possible in the system design process. This way, an inappropriate poor investment of money, time and effort can be prevented. Here also lies an opportunity for researchers to map the types of tasks for which personalization can have added value for each kind of system (e.g., a personalized Internet search or personalized tourist guides).

## 6.4 On the Importance of the User Experience

Not all features of a personalized system are equally important. Zhang and Von Dran (2001) make a distinction between three types: basic, perform-

ance and exciting features. Basic features support the minimal needs of users and are often taken for granted by users. Performance features cater to the consciously stated needs of users and their implementation to ensure that a system becomes a viable competitor in the market. Exciting features, finally, exceed user expectations and foster user loyalty. Over time, user expectations shift and features that were initially exciting become performance features, which in turn may later become basic features.

As the literature review reported in chapter 4 indicated, most research on personalization focuses on optimizing the effectiveness and efficiency of technology and, thus, on optimizing basic features. The user experience of personalization, which is comprised of factors like trust, predictability, and so on, has received relatively little attention from the research community. Díaz, García and Gervás (2008) have strikingly articulated the limited value of this focus in an article that compares system-centered evaluations (which focuses on effectiveness and efficiency) and user-centered evaluations of personalization (which focuses on user perceptions). They state:

"[System-centered evaluation] may be a good alternative for guiding the development of algorithms, but it is a poor approach for the guiding the development of systems designed for human users. The motor car industry has long since accepted that the refinements and tuning possibilities that can make a Formula (1) prototype maximally efficient at the racing track need not be the kind of feature that a user wants in the car he drives to work every day, irrespective of the objective data presented in the dashboard or provided by the stopwatch." (Díaz, García, & Gervás, 2008, pp. 1303-1304)

Usability and the user experience, which are underrepresented in personalization research, may be reflected in performance features, but they may also in part comprise basic features. In chapter 2, for example, we showed that trust in the technology and especially perceived controllability are very important aspects of the interaction between a user and a personalized system. A lack of trust in the technology or a lack of tools for making personalization seem controllable may lead a user to decide not to use the technology, as his or her minimal needs may not be adequately addressed.

Failing to provide performance features in personalized systems is a risky course of action for organizations that aim to compete with other players in the market. Their systems may function well, but they do not offer users that extra amount necessary to make them want to use a particular system instead of one provided by a competitor. Implementing performance

features by taking into account usability and the user experience in the design of personalization is a strategy that is likely to lead to personalized systems that not only do what they are supposed to do but are also attractive to (prospective) users.

The heavy research focus on effectiveness and efficiency has resulted in a limited understanding of the interaction between the user and the personalized system. Consequently, researchers have not been able to adequately provide designers with the insights or tools they need to create personalized systems that go beyond providing effective technology. The research community thus must broaden its focus and pay attention to not only effectiveness and efficiency but also to the user experience of personalization to cater to all basic user needs and to contribute to the creation of personalized systems that can compete in the market.

## 6.5 Closing Remarks

In this thesis, I have taken a user-centered approach to design and evaluation of personalization. The four studies reported in this thesis have expanded the toolkits of designers and evaluators alike by yielding concrete design guidelines (specifically on the role of trust and controllability) and by uncovering the value of different methods for the requirements engineering phase as well as the formative evaluation phase in the UCD process for personalization.

I hope this thesis will inspire researchers and developers to apply a user-centered approach when developing personalized features and to continue to investigate how the UCD process can be optimized to adequately address personalization. In the end, such efforts serve one goal: to create technology that serves people and society.
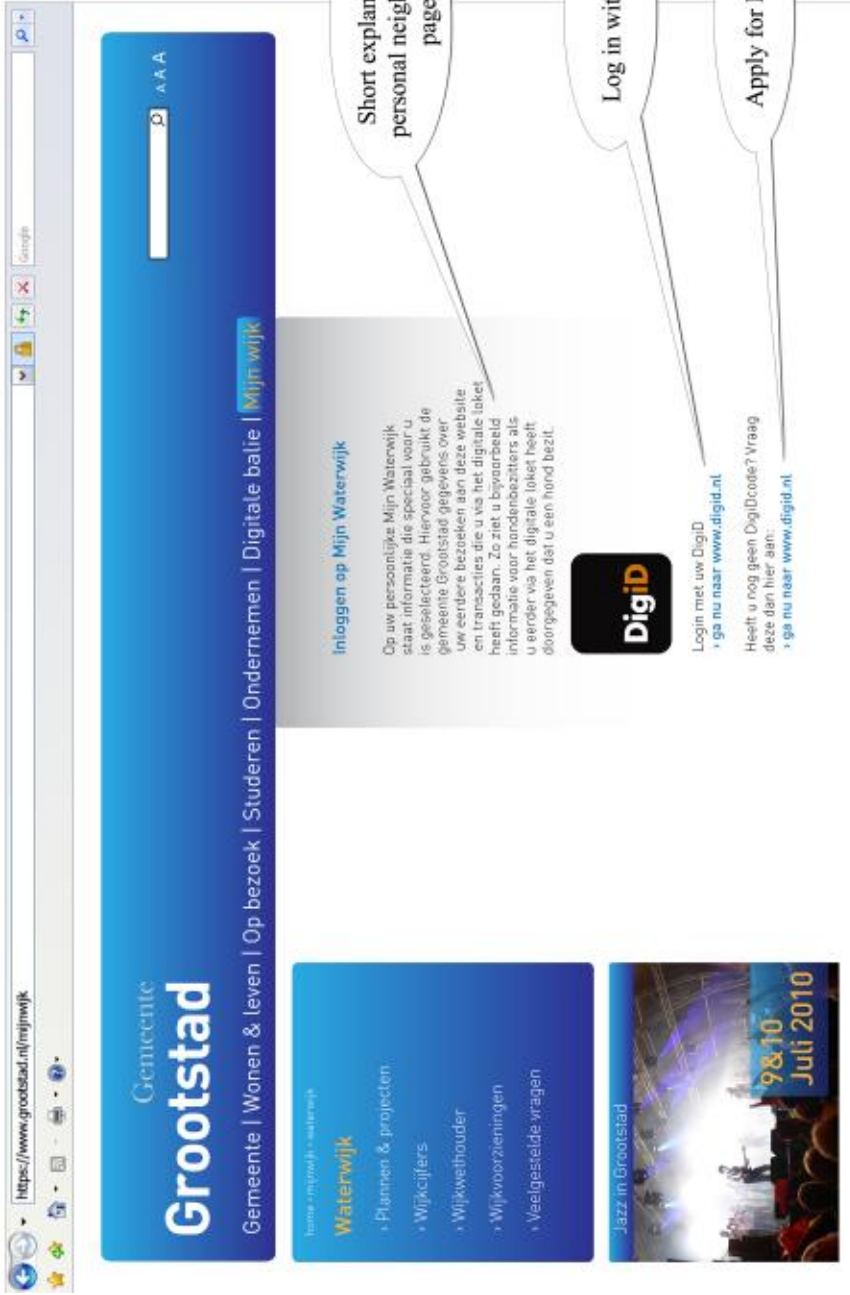
# Appendix Chapter 2

# Appendix A: Exemplary screenshots
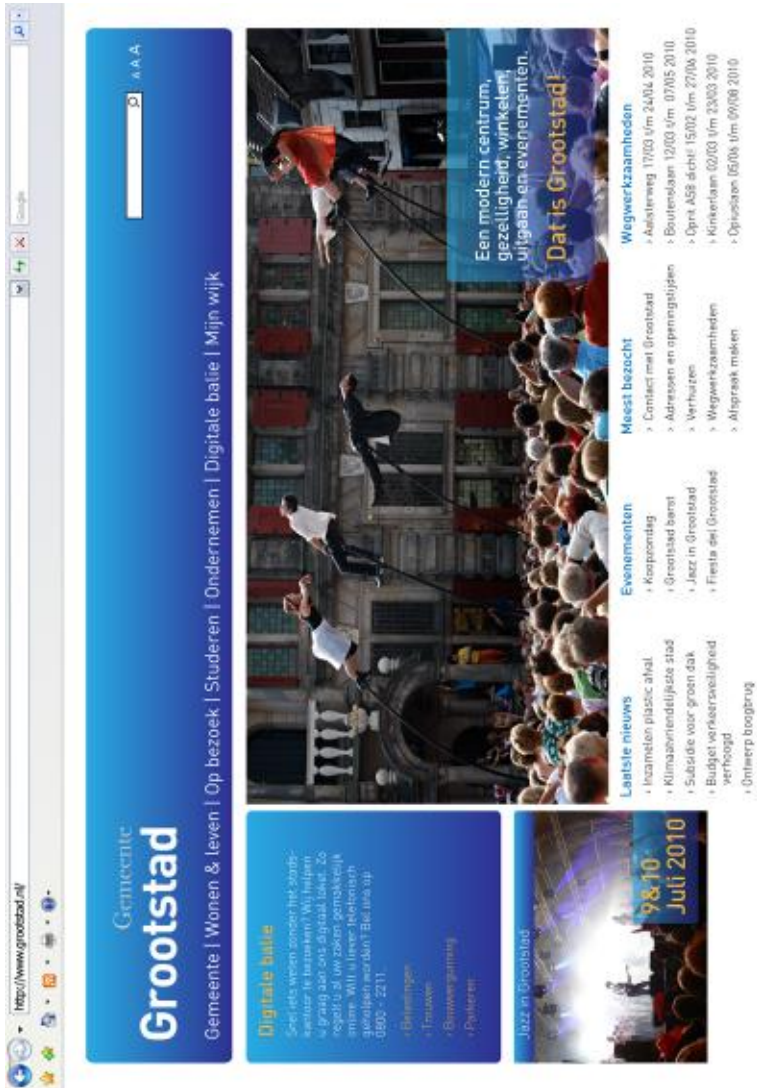


Grootstad homepage
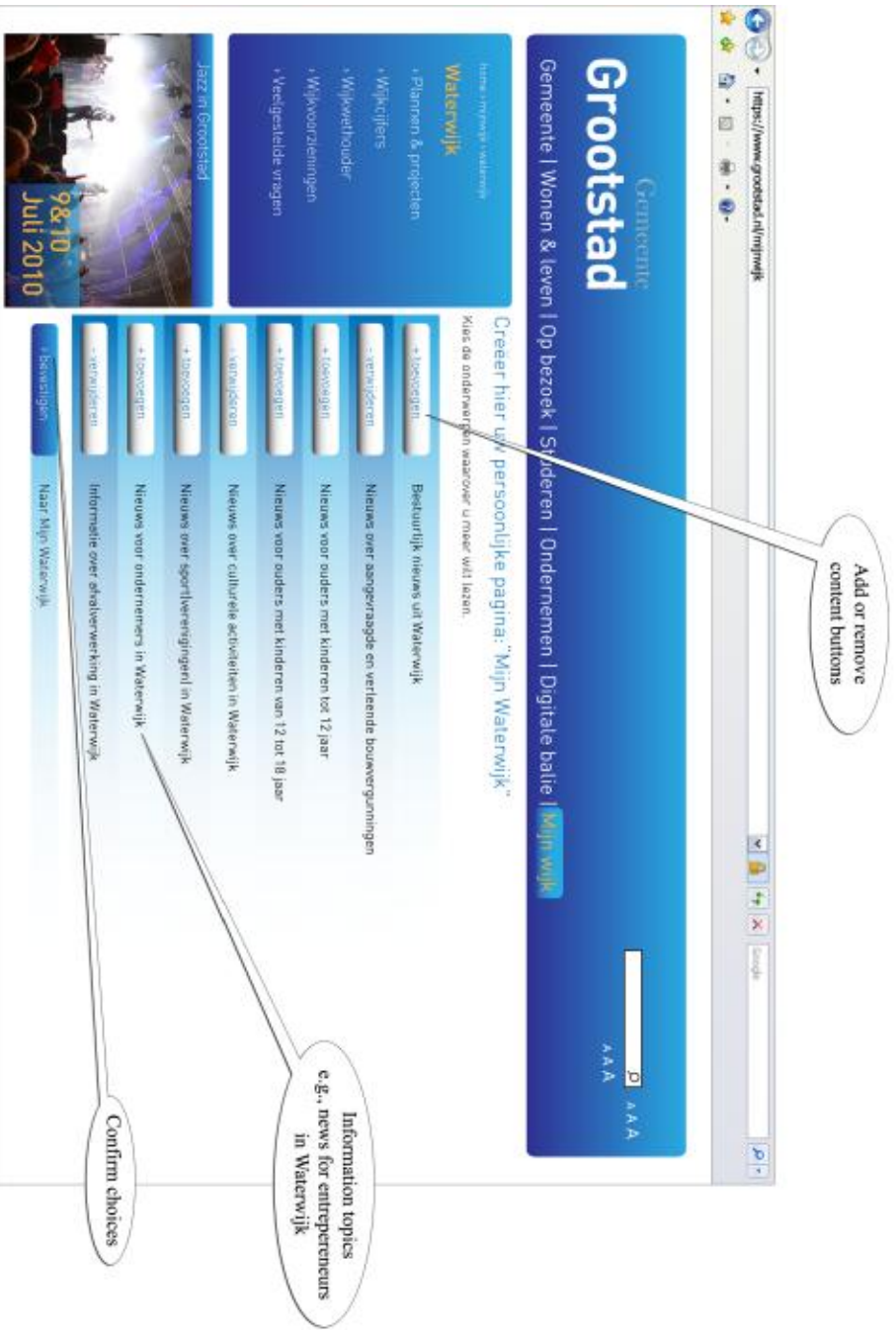
Customizing your personal neighborhood page (adaptable condition)

Short explanation of personal neighborhood page

Log in with DigiD

Apply for DigiD

Log-in screen for the adaptive/behavior condition

Personal neighborhood page for Karin van Heek (adaptive/demographic condition)

**Issued building permits within 500 metres of your house**

**Information for people with a high income**

**Information for parents with young children**

**Information for home owners: subsidy for solar panels**

Gemeente

**Grootstad**

Gemeente | Wonen & leven | Op bezoek | Studeren | Ondernemen | Digitale balie | Mijn wijk

Mijn Waterwijk van Karin van Heek

Waterwijk
- Plannen & projecten
- Wijkcijfers
- Wijkwethouder
- Wijkvoorzieningen
- Veelgestelde vragen

Jazz in Grootstad
9 & 10 Juli 2010

Bouwvergunningen binnen 500m van uw huis

Informatie voor ouders met jonge kinderen

Informatie voor mensen met een hoog inkomen

Subsidie aanschaf zonnepanelen

# Appendix B: Survey items

| Construct | | Item | Source |
|---|---|---|---|
| Trust in municipality | | | Bélanger & Carter, 2008 |
| | 1 | I can trust my municipality. | |
| | 2 | My municipality handles my personal data carefully. | |
| | 3 | My municipality has my best interests in mind. | |
| | 4 | My municipality is not trustworthy. | |
| Trust in technology | | | McKnight, Choudhury & Kacmar, 2002 |
| | 1 | The security on this page does not set my mind at rest. | |
| | 2 | The law and security technology protect me well against problems on this page. | |
| | 3 | Your personal data are protected well when you use this page. | |
| | 4 | This page is not safe. | |
| Perceived controllability | | | Liu, 2003 |
| | 1 | I have a lot of control over what I can do on this page. | |
| | 2 | On this page, you can choose freely what you want to see. | |
| | 3 | On this page, you have absolutely no control over what you will see. | |
| | 4 | I can determine for myself what happens on this page. | |
| Intention to use | | | |
| | 1 | If this page existed for my neighborhood, I would use it. | Venkatesh & Davis, 2000 |
| | 2 | If this page existed for my neighborhood, I would definitely use it. | Venkatesh & Davis, 2000 |
| | 3 | I would recommend such a page to others. | Gefen, Karahanna & Straub, 2003 |
| | 4 | I hope my municipality makes one of these pages for my neighborhood. | |

# Appendix C: Initial item factor loadings and Cronbach's alphas

|  | Baseline | Adaptable | Adaptive/ demographic | Adaptive/ behavior | Adaptive/ psycho-graphic |
|---|---|---|---|---|---|
| Trust in organization | alpha .87 | alpha .85 | alpha .89 | alpha .90 | alpha .93 |
| TO1 | .80 | .74 | .82 | .80 | .85 |
| TO2 | .65 | .72 | .78 | .81 | .81 |
| TO3 | .79 | .75 | .76 | .83 | .81 |
| TO4 | .66 | .60 | .68 | .69 | item removed |
| Trust in technology | alpha .76 | alpha .85 | alpha .83 | alpha .88 | alpha .82 |
| TT1 | item removed | .59 | .58 | .63 | .55 |
| TT2 | .60 | .73 | .67 | .81 | .66 |
| TT3 | .48 | .79 | .73 | .75 | .77 |
| TT4 | .60 | .66 | .68 | .75 | .61 |
| Perceived controllability | alpha .74 | alpha .75 | alpha .76 | alpha .82 | alpha .84 |
| PC1 | .59 | .56 | .62 | .68 | .68 |
| PC2 | .59 | .55 | .51 | .69 | .72 |
| PC3 | .41 | item removed | item removed | .53 | .59 |
| PC4 | .55 | .58 | .60 | .67 | .73 |
| Intention to use | alpha .89 | alpha .93 | alpha .89 | alpha .94 | alpha .95 |
| IU1 | .72 | .84 | .75 | .84 | .88 |
| IU2 | .80 | .87 | .80 | .89 | .93 |
| IU3 | .72 | .78 | .73 | .83 | .85 |
| IU4 | .77 | .85 | .78 | .86 | .89 |

# Appendix Chapter 4

## Reports included in the review

Alpert, S. R., Karat, J., Karat, C. M., Brodie, C. & Vergo, J. G. (2003). User attitudes regarding a user-adaptive eCommerce web site. *User Modeling and User-Adapted Interaction, 13*(4), 373-396.

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

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

Biemans, M. C. M., Van Kranenburg, H. & Lankhorst, M. (2001). *User evaluations to guide the design of an extended personal service environment for mobile services*. Paper presented at the 5th international symposium on wearable computers, October 7-9, Zurich, Switzerland.

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

Bohnenberger, T., Jameson, A., Krüger, A., & Butz, A. (2002). Location-aware shopping assistance: evaluation of a decision-theoretic approach. In F. Paternò (Ed.), *Human computer interaction with mobile devices* (pp. 155-169). Berlin: Springer.

Brusilovsky, P. & Anderson, J. (1998). *Act-r electronic bookshelf: An adaptive system to support learning act-r on the web*. Paper presented at Webnet 98 world conference of the WWW, internet, and intranet, November 7-12, Orlando, FL, USA.

Cawsey, A. J., Jones, R. B., & Pearson, J. (2000). The evaluation of a personalised health information system for patients with cancer. *User modeling and user-adapted interaction, 10*(1), 47-72.

Chesnais, P. R., Mucklo, M. J., & Sheena, J. A. (1995). *The Fishwrap personalized news system*. Paper presented at the second international workshop on community networking 'Integrated multimedia services to the home', June, 20-22, Princeton, NJ, USA.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring issues of user model transparency and proactive behavior in an office environment control system. *User modeling and user-adapted interaction, 15*(3/4), 235-273.

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

Conlan, O., O'Keeffe, I. & Tallon, S. (2006). Combining adaptive hypermedia techniques and ontology reasoning to produce dynamic personalized news services. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 81-90). Berlin: Springer.

Conlan, O. & Wade, V. P. (2004). Evaluation of APeLS - an adaptive elearning service based on the multi-model, metadata-driven approach. In P. de Bra & W. Nejdl (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp 291-295).

Cox, R., O'Donnel, M. & Oberlander, J. (1999). *Dynamic versus static hypermedia in museum education: An evaluation of ilex, the intelligent labelling explorer*. Paper presented at the artificial intelligence in education conference, July 19-23, Le Mans, France.

De Almeida, P. & Yokoi, S. (2003). Interactive character as a virtual tour guide to an online museum exhibition. In D. Bearman and J. Trant (Eds.), *Museums and the web 2003: selected papers from an international conference* (pp. 191-198). Pittsburgh: Archives & museum informatics.

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

Díaz, A., Gervás, P. & García, A. (2005). Evaluation of a system for personalized summarization of web contents. In L. Ardissono, P. Brna and A. Mitrovic (Eds.), *User modeling 2005* (pp. 453-462). Berlin: Springer.

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

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

Field, A., Hartel, P., & Mooij, W. (2001). *Personal DJ, an open architecture for personalised content delivery*. Paper presented at the the tenth international World Wide Web conference, May, 1-5, Hong Kong.

Gates, K. F., Lawhead, P. B., & Wilkings, D. E. (1998). Toward an adaptive WWW: A case study in customised hypermedia. *New review of hypermedia and multimedia, 4*(1), 89-113.

Gena, C. (2002). *An empirical evaluation of an adaptive web site. Paper presented at the seventh international conference on intelligent user interfaces, January 13-16,*San Francisco, CA, USA.

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

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

Goren-Bar, D. & Kuflik, T. (2004). *Don't miss-r: recommending restaurants through an adaptive mobile system*. Paper presented at the ninth international conference on intelligent user interfaces, January 13-16 Funchal, Portugal.

Graziola, I., Pianesi, F., Zancanaro, M. & Goren-Bar, D. (2005). *Dimensions of adaptivity in mobile systems: personality and people's attitudes*. Paper presented at the tenth international conference on intelligent user interfaces, January 9-12, San Diego, CA, USA.

Gregor, P., Dickinson, A., Macaffer, A., & Andreasen, P. (2003). Seeword: A personal word processing environment for dyslexic computer users. *British journal of educational technology, 34*(3), 341-355.

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

Henderson, R., Rickwood, D., & Roberts, P. (1998). The beta test of an electronic supermarket. *Interacting with computers, 10*(4), 385-399.

Höök, K. (1997). *Evaluating the utility and usability of an adaptive hypermedia system*. Paper presented at the 2nd international conference on Intelligent user interfaces, January 6-9, Orlando, FL, USA.

Hyldegaard, J., & Seiden, P. (2004). My E-journal: Exploring the usefulness of personalized access to scholarly articles and services [Electronic Version]. *Information research*, 9, paper 181 from
http://informationr.net/ir/9-3/paper181.html.

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

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

Järvinen, T., (2005). *Hybridmedia as a tool to deliver personalised product-specific information about food. Report of the TIVIK project*. Helsinki: VTT publications.

Kaasinen, E. (2003). User needs for location aware mobile services. *Personal ubiquitous computing, 7*(1), 70-79.

Karat, C., Brodie, C., Karat, J., Vergo, J., & Alpert, S. R. (2003). Personalizing the User Experience on ibm.com. *IBM systems journal, 42*(4), 686-701.

Kavcic, A., Privosnik, M., Marolt, M. & Divjak, S. (2002). *Educational hypermedia: An evaluation study*. Paper presented at the the mediterranean electrotechnical conference, May 7-9, Cairo, Egypt.

Ketamo, H. (2003). Xtask: An adaptable learning environment. *Journal of computer assisted learning, 19*(3), 360-370.

Kiris, E. (2004). *User-centered eservice design and redesign*. Paper presented at the conference on human factors in computing systems, April 24-29, Vienna, Austria.

Kolari, J., Laakko, T., Hiltunen, T., Ikonen, V., Kulju, M., Suihkonen, R., et al. (2004). *Context-aware services for mobile users (technology and user experiences)*. VTT technical research centre of Finland.

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

Kuiper, P. M., Van Dijk, E. M. A. G. & Boerma, A. K. (2006). *Adaptive municipal e-forms*. Paper presented at the fifth workshop on user-centred design and evaluation of adaptive systems, held in conjunction with the 4th international conference on adaptive hypermedia & adaptive web based systems, June 20, Dublin, Ireland.

Masthoff, J. (2006). *The user as wizard: A method for early involvement in the design and evaluation of adaptive systems*. Paper presented at the fifth workshop on user-centred design and evaluation of adaptive systems, held in conjunction with the 4th international conference on adaptive hypermedia & adaptive web based systems, June 20, Dublin, Ireland.

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

Muntean, C. H. & McManis, J. (2006). The value of QoE-based adaptation approach in educational hypermedia: Empirical evaluation. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 121-130). Berlin: Springer.

O'Grady, M. J. & O'Hare, G. M. P. (2004). *Enabling customized & personalized interfaces in mobile computing*. Paper presented at the ninth international conference on intelligent user interfaces, January 13-16 Funchal, Portugal.

Patel, D. & Marsden, G. (2004). *Customizing digital libraries for small screen devices*. Paper presented at the annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries, Stellenbosch, Western Cape, South Africa.

Pateli, A. G., Giaglis, G. M., & Spinellis, D. D. (2005). Trial evaluation of wireless info-communication and indoor location-based services in exhibition shows. In P. Bozanis & E. N. Houstis (Eds.), *Advances in informatics* (pp. 199-210). Berlin: Springer.

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

Romero, C., Ventura, S., Hervas, C., & De Bra, P. (2006). An authoring tool for building both mobile adaptable tests and web-based adaptive or classic tests. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 203-212). Berlin: Springer.

Schmidt-Belz, B., & Posland, S. (2003). *User validation of a mobile tourism services*. Paper presented at the Workshop on HCI mobile guides, held in conjunction with MobileHCI03, September, 8-11, Udine, Italy.

Smith, H., Fitzpatrick, G., & Rogers, Y. (2004). *Eliciting reactive and reflective feedback for a social communication tool: A multi-session approach*. Paper presented at the 5th conference on Designing interactive systems: processes, practices, methods, and techniques August, 1-4, Cambridge, MA, USA.

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

Södergard, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Järvinen, T., Kaasinen, E., et al. (1999). Integrated multimedia publishing: combining TV and newspaper content on personal channels. *Computer networks, 31*(11-16), 1111-1128.

Stary, C., & Totter, A. (2003). Measuring the adaptability of universal accessible systems. *Behavior & information technology, 22*(2), 101-116.

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

Stein, A. (1997). Usability and assessments of multimodal interaction in the SPEAK! system: An experimental case study. *The new review of hypermedia and multimedia, 3*(1), 159-180.

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

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

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

Weibelzahl, S. (2003). *Evaluation of adaptive systems*. Freiburg: Pedagogical university Freiburg.

Weibelzahl, S., Jedlitschka, A. & Ayari, B. (2006). *Eliciting requirements for an adaptive decision support system through structured user interviews*. Paper presented at the fifth workshop on user-centred design and evaluation of adaptive systems, held in conjunction with the 4th international conference on adaptive hypermedia & adaptive web based systems, June 20, Dublin, Ireland.

Yue, W., Mu, S., Wang, H. & Wang, G. (2005). *TGH: A case study of designing natural interaction for mobile guide systems*. Paper presented at the 7th international conference on human computer interaction with mobile devices & services, September 19-22, Salzburg, Austria.

# Appendix Chapter 5

## Questionnaire and interview items*

- Do you have the feeling that Prospector works predictably? If so, why?
- Do you have the feeling that you understand how Prospector works? If so, why?
- Are there certain parts of Prospector you think are hard to understand? And, if so, which ones are they? And why do you think these are hard to understand?
- Do you have the feeling you give away control when you use Prospector? If so, why?
- Do you have the feeling that giving Prospector information about yourself (like indicating your interests or rating search results) costs too much time and effort? If so, why?
- Do you feel that Prospector infringes on your privacy? If so, why?
- Do you feel that gearing search results to your personal situation goes at the expense of discovering new (kinds of) information? If so, why?
- Do you feel that Prospector is good enough to gear search results to your personal situation?
- What reasons would prompt you to use - or not use - Prospector?
- Do you think Prospector is better than Google? If so, why?

*Items are translated from Dutch

# References

Ackerman, M. S., Cranor, L. F., & Reagle, J. (1999). *Privacy in e-commerce: examining user scenarios and privacy preferences*. Paper presented at the 1st ACM conference on electronic commerce, November 3-5, Denver, CO, USA.

Akoumianakis, D., Grammenos, D., & Stephanidis, C. (2001). User interface adaptation: evaluation perspectives. In C. Stephanidis (Ed.), *User interfaces for all. Concepts, methods, and tools* (pp. 339-352). Mahwah: Lawrence Erlbaum.

Allwood, C. M., & Kalén, T. (1997). Evaluating and improving the usability of a user manual. *Behavior & information technology, 16*(1), 43-57.

Alpert, S. R., & Vergo, J. G. (2007). User-centered evaluation of personalized web sites: what's unique? In P. Zaphiris & S. Kurniawan (Eds.), *Human computer interaction. Research in web design and evaluation* (pp. 257-272). Hershey: Idea group publishing.

Ambrosini, V., & Bowman, C. (2001). Tacit knowledge: some suggestions for operationalization. *Journal of management studies, 38*(6), 811-829.

Ammenwerth, E., Iller, C., & Mansmann, U. (2003). Can evaluation studies benefit from triangulation? A case study. *International journal of medical informatics, 70*(2/3), 237-248.

Andersen, K. V., & Henriksen, H. Z. (2006). E-government maturity models: extension of the Layne and Lee model. *Government information quarterly, 23*(2), 236-248.

Andrade, E. B., Kaltcheva, V., & Weitz, B. (2002). Self-disclosure on the web: The impact of privacy policy, reward, and company reputation. *Advances in consumer research, 29*(1), 350-353.

Awad, N. F., & Krishnan, M. S. (2006). The personalization privacy paradox: an empirical evaluation of information transparency and the willingness to be profiled online for personalization. *MIS Quarterly, 30*(1), 13-28.

Baaren, E., Van de Wijngaert, L., & Huizer, E. (2008). *Not another TAM paper: relating individual and context characteristics to the adoption of HDTV*. Paper presented at the international communication association conference, May 22-26, Montreal, Canada.

Barkhuus, L., & Dey, A. (2003). Is context-aware computing taking control away from the user? Three levels of interactivity examined. In A. Dey, A. Schmidt & J. F. McCarthy (Eds.), *UbiComp 2003: Ubiquitous Computing* (pp. 149-156). Heidelberg: Springer.

Beane, T. P., & Ennis, D. M. (1987). Market segmentation: A review. *European journal of marketing, 21*(5), 20-42.

Becker, S. A. (2004). E-government visual accessibility for older adult users. *Social science computer review, 22*(1), 11-23.

Bélanger, F., & Carter, L. (2008). Trust and risk in e-government adoption. *Journal of strategic information systems, 17*(2), 165-176.

Beldad, A., De Jong, M., & Steehouder, M. (2010). How shall I trust the faceless and the intangible? A literature review on the antecedents of online trust. *Computers in human behavior, 26*(5), 857-869.

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

Benchmark personalization of governmental eServices for citizens. (2008). Retrieved January 7, 2009, from http://www.e-overheid.nl/e-overheid-2.0/live/binaries/pip/bestanden/benchmark-report—definitief_4aug.pdf.

Benyon-Davies, P., Tudhope, D., & Mackay, H. (1999). Information systems prototyping in practice. *Journal of information technology, 14*(1), 107-120.

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

Bohnenberger, T., Jameson, A., Krüger, A., & Butz, A. (2002). Location-aware shopping assistance: evaluation of a decision-theoretic approach. In F. Paternò (Ed.), *Human computer interaction with mobile devices* (pp. 155-169). Berlin: Springer.

Bradburn, N. M., Sudman, S., & Wansink, B. (2004). *Asking questions*. San Francisco: Jossey-Bass.

Brennan, R., Baines, P., & Garneau, P. (2003). *Contemporary strategic marketing*. Houndmills: Palgrave Macmillan.

Bruns, A. (2007). *Produsage: Towards a broader framework for user-led content creation*. Paper presented at Creativity & cognition, June 13-15, Washington, DC, USA.

Brusilovsky, P. (1996). Methods and techniques of adaptive hypermedia. *User modeling and user-adapted interaction, 6*(2/3), 87-129.

Brusilovsky, P. (2001). Adaptive hypermedia. *User modeling and user-adapted interaction, 11*(1/2), 87-110.

Brusilovsky, P., Karagiannidis, C., & Sampson, D. (2001). *Benefits of layered evaluation of adaptive applications and services*. Paper presented at the empirical evaluation of adaptive systems workshop, held in conjunction with the 8th international conference on user modeling, July 13, 2001, Sonthofen, Germany.

Buchauer, A., Pohl, U., Kurzel, N., & Haux, R. (1999). Mobilizing a health professional's workstation: Results of an evaluation study. *International journal of medical informatics, 54*(2), 105-114.

Bunt, A., McGrenere, J., & Conati, C. (2007). Understanding the utility of rationale in a mixed-initiative system for GUI customization. In C. Conati, K. McCoy & G. Paliouras (Eds.), *User modeling 2007* (pp. 147-156). Berlin: Springer.

Byrt, T. (1996). How good is that agreement? *Epidemiology, 7*(5), 561.

Canny, J. (2006). The future of human-computer interaction. *ACM Queue, 4*(6), 24-32.

Carmagnola, F., Cena, F., Console, L., Cortassa, O., Gena, C., Goy, A., et al. (2008). Tag-based user modeling for social multi-device adaptive guides. *User modeling and user-adapted interaction, 18*(5), 497-538.

Carrol, C., Marsden, P., Soden, P., Naylor, E., New, J., & Dornan, T. (2002). Involving users in the design and usability evaluation of a clinical decision support system. *Computer methods and programs in biomedicine, 69*(2), 123-135.

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: citizen trust, innovation and acceptance factors. *Information systems journal, 15*(1), 5-25.

Carter, P. (2007). Liberating usability testing. *Interactions, 14*(2), 18-22.

Cawsey, A. J., Grasso, F., & Paris, C. (2007). Adaptive information for consumers of healthcare. In P. Brusilovsky, A. Kobsa & W. Nejdl (Eds.), *The adaptive web* (pp. 465-484). Heidelberg: Springer.

Cawsey, A. J., Jones, R. B., & Pearson, J. (2000). The evaluation of a personalised health information system for patients with cancer. *User modeling and user-adapted interaction, 10*(1), 47-72.

Chellappa, R. K., & Sin, R. G. (2005). Personalization versus privacy: An empirical examination of the online consumer's dilemma. *Information technology and management, 6*(2/3), 181-202.

Chesnais, P. R., Mucklo, M. J., & Sheena, J. A. (1995). *The Fishwrap personalized news system*. Paper presented at the 2nd international workshop on community networking 'Integrated multimedia services to the home', June 20-22, Princeton, NJ, USA.

Cheverst, K., Byun, H. E., Fitton, D., Sas, C., Kray, C., & Villar, N. (2005). Exploring issues of user model transparency and proactive behavior in an office environment control system. *User modeling and user-adapted interaction, 15*(3/4), 235-273.

Chin, D. N. (2001). Empirical evaluation of user models and user-adapted systems. *User modeling and user-adapted interaction, 11*(1/2), 181-194.

Coble, J. M., Karat, J., & Kahn, M. G. (1997). *Maintaining a focus on user requirements throughout the development of clinical workstation soft-*

*ware*. Paper presented at the SIGCHI conference on human factors in computing systems, May 22-27, Atlanta, USA.

Colineau, N., & Paris, C. (2009). Does tailoring help people find the information they need? *New review of hypermedia and multimedia, 15*(3), 267-286.

Cooper, A. (1999). *The inmates are running the asylum. Why high-tech products drive us crazy and how to restore the sanity*. Indianapolis: SAMS.

Cordes, R. E. (2001). Task-selection bias: A case for user-defined tasks. *International journal of human-computer interaction, 13*(4), 411-419.

Corritore, C. L., Kracher, B., & Wiedenbeck, S. (2003). On-line trust: concepts, evolving themes, a model. *International journal of human-computer studies, 58*(6), 737-758.

Cover, R. (2006). Audience inter/active: Interactive media, narrative control and reconceiving audience history. *New media & society, 8*(1), 139-158.

Cramer, H., Evers, V., Ramlal, S., Van Someren, M., Rutledge, L., Natalia, S., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User modeling and user-adapted interaction, 18*(5), 455-496.

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

Culnan, M. T. (1993). "How did they get my name?": An exploratory investigation of consumer attitudes toward secondary information use. *MIS Quarterly, 17*(3), 341-363.

Damodaran, L. (1996). User involvement in the systems design process - a practical guide for users. *Behavior & information technology, 15*(6), 363-377.

Davis, B. G. (1982). Strategies for information requirements determination. *IBM systems journal, 21*(1), 4-30.

Davis, F. D. (1986). *A technology acceptance model for empirically testing new end-user information systems: theory and results*. Massachusetts institute of technology, Massachusetts.

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Quarterly, 13*(3), 319-340.

Davis, F. D., & Venkatesh, V. (2004). Toward preprototype user acceptance testing of new information systems: implications for software project management. *IEEE transactions on engineering management, 51*(1), 31-46.

De Jong, M., & Schellens, P. J. (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliabil-

ity of evaluation methods? *IEEE transactions on professional communication, 43*(3), 242-260.

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

Díaz, A., Gercía, A., & Gervás, P. (2008). User-centred versus system-centred evaluation of a personalization system. *Information processing and management, 44*(3), 1293-1307.

Dicks, R. S. (2002). *Mis-usability: On the uses and misuses of usability testing*. Paper presented at the 20th annual international conference on computer documentation, October 20-23, Toronto, Canada.

Donker, A., & Markopoulos, P. (2002). A comparison of think-aloud, questionnaires and interviews for testing usability with children. In X. Faulkner, J. Finlay & F. Detienne (Eds.), *Proceedings of human computer interaction 2002* (pp. 305-316). London: Springer.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). *A comparison of usability techniques for evaluating design*. Paper presented at the 2nd conference on designing interactive systems: processes, practices, methods, and techniques, August 18-20, Amsterdam, the Netherlands.

Duh, H. B., Tan, G. C. B., & Chen, V. H. (2006). *Usability evaluation for mobile devices: A comparison of laboratory and field tests*. Paper presented at the 8th conference on human-computer interaction with mobile devices and services, September 12-15, Helsinki, Finland.

Ebling, M. R., & John, B. E. (2000). *On the contributions of different empirical data in usability testing*. Paper presented at the 3rd conference on designing interactive systems: processes, practices, methods, and techniques, August 17-19, Brooklyn, NY, USA.

Egger, F. N. (2003). *From interactions to transactions: Designing the trust experience for business-to-consumer electronic commerce.* Technical University of Eindhoven, Eindhoven, the Netherlands.

Field, A., Hartel, P., & Mooij, W. (2001). *Personal DJ, an open architecture for personalised content delivery*. Paper presented at the 10th international World Wide Web conference, May 1-5, Hong Kong.

Følstad, A., Jørgensen, H. D., & Krogstie, J. (2004). *User Involvement in e-Government Development Projects*. Paper presented at NordCHI'04, Tampere, Finland.

Fossey, E., Harvey, C., McDermott, F., & Davidson, L. (2002). Understanding and evaluating qualitative research. *Australian & New Zealand journal of psychiatry, 36*(6), 717-733.

Fowler Jr., F. J. (1995). *Improving survey questions. Design and evaluation*. Thousand oaks: Sage.

Gabrielli, S., & Jameson, A. (2009). *Differences and changes in preferences regarding personalized systems: A user-centred design perspective*. Paper presented at the 6th workhop on user-centred design and evaluation of adaptive systems, held in conjunction with the international conference on user modeling, adaptation and personalization, June 26, Trento, Italy.

Gates, K. F., Lawhead, P. B., & Wilkings, D. E. (1998). Toward an adaptive WWW: A case study in customised hypermedia. *New review of hypermedia and multimedia, 4*(1), 89-113.

Gefen, D., Karahanna, E., & Straub, D. W. (2003). Trust and TAM in online shopping: An integrated model. *MIS Quarterly, 27*(1), 51-90.

Gena, C. (2005). Methods and techniques for the evaluation of user-adaptive systems. *The knowledge engineering review, 20*(1), 1-37.

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

Gena, C., & Weibelzahl, S. (2007). Usability engineering for the adaptive web. In P. Brusilovsky, A. Kobsa & W. Nejdl (Eds.), *The adaptive web* (pp. 720-762). Berlin: Springer.

Godek, J., & Yates, J. F. (2005). Marketing to individual consumers online: The influence of perceived control. In C. P. Haugtvedt, K. A. Machleit & R. F. Yalch (Eds.), *Online consumer psychology. Understanding and influencing consumer behavior in the virtual world* (pp. 225-244). Mahwah, NJ: Lawrence Erlbaum associates.

Goodhue, D. L., & Thompson, R. L. (1995). Task-technology fit and individual performance. *MIS Quarterly, 19*(2), 213-236.

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

Gould, J. D., Boies, S. J., & Lewis, C. (1991). Making usable, useful, productivity. Enhancing computer applications. *Communication of the ACM, 34*(1), 74-85.

Gould, J. D., & Lewis, C. (1985). Designing for usability: key principles and what designers think. *Communications of the ACM, 28*(3), 300-311.

Grabner-Krauter, S. (2002). The role of consumers' trust in online-shopping. *Journal of business ethics, 39*(1/2), 43-50.

Grady, H. M. (2000). *Web site design: a case study in usability testing using paper prototypes*. Paper presented at the IEEE professional communication conference, September 24-27, Cambridge, USA.

Graeff, T. R., & Harmon, S. (2002). Collecting and using personal data: consumers' awareness and concerns. *Journal of consumer marketing, 19*(4), 302-318.

Gray, W. D., & Salzman, M. C. (1998). Damaged merchandise? A review of experiments that compare usability evaluation methods. *Human-computer interaction, 13*(3), 203-261.

Gregor, P., Dickinson, A., Macaffer, A., & Andreasen, P. (2003). Seeword: A personal word processing environment for dyslexic computer users. *British journal of educational technology, 34*(3), 341-355.

Gremler, D. D. (2004). The critical incident technique in service research. *Journal of service research, 7*(1), 65-89.

Gulliksen, J., Göransson, B., Boivie, I., Blomkvist, S., Persson, J., & Cajander, Å. (2003). Key principles for user-centred systems design. *Behavior & information technology, 22*(6), 397-409.

Haley, R. I. (1968). Benefit segmentation: A decision-oriented research tool. *Journal of marketing, 32*(3), 30-35.

Haraldsen, M., Stray, T. D., Päivärinta, T., & Sein, M. K. (2004). *Developing e-Government portals: from life-events through genres to requirements*. Paper presented at the 11th Norwegian conference on information systems, November 29–December 1, Stavanger, Norway.

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International journal of human-computer interaction, 13*(4), 373-410.

Henderson, R., Rickwood, D., & Roberts, P. (1998). The beta test of an electronic supermarket. *Interacting with computers, 10*(4), 385-399.

Henderson, R. D., Smith, M. C., Podd, J., & Varela-Alvarez, H. (1995). A comparison of the four prominent user-based methods for evaluating the usability of computer software. *Ergonomics, 38*(10), 2030-2044.

Herder, E. (2006). *Forward, back and home again. Analyzing user behavior on the web*. Enschede: University of Twente.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: Does thinking aloud affect behavior and mental workload? *Behavior & information technology, 28*(2), 165-181.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: a chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 421-443.

Høegh, R. T., & Jensen, J. J. (2008). A case study of three software projects: Can software developers anticipate the usability problems in their software? *Behavior & information technology, 27*(4), 307-312.

Hoek, J., Gendall, P., & Esslemont, D. (1996). Market segmentation: A search for the Holy Grail? *Journal of marketing practice, 2*(1), 25-34.

Holzinger, A. (2005). Usability engineering methods for software developers. *Communications of the ACM, 49*(1), 71-74.

Höök, K. (1997). *Evaluating the utility and usability of an adaptive hypermedia system*. Paper presented at the 2$^{nd}$ international conference on intelligent user interfaces, January 6-9, Orlando, FL, USA.

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

Hoppmann, T. K. (2009). Examining the 'point of frustration'. The think-aloud method applied to online search tasks. *Quality and quantity, 43*(2), 211-224.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behavior & information technology, 29*(1), 97-111.

Hornbæk, K., & Frøkjær, E. (2005). *Comparing usability problems and redesign proposals as input to practical systems development*. Paper presented at the SIGCHI conference on human factors in computing systems, April 2-7, Portland, OR, USA.

Horst, M., Kuttschreuter, M., & Gutteling, J. (2007). Perceived usefulness, personal experiences, risk perception and trust as determinants of adoption of e-government services in The Netherlands. *Computers in human behavior, 23*(4), 1838-1852.

Hu, R., & Pu, P. (2010). A study on user perception of personality-based recommender systems. In P. De Bra, A. Kobsa & D. Chin (Eds.), *User modeling, adaptation, and personalization* (pp. 291-302). Berlin: Springer.

Hyldegaard, J., & Seiden, P. (2004). My E-journal: Exploring the usefulness of personalized access to scholarly articles and services [Electronic Version]. *Information research*, 9, paper 181 from http://informationr.net/ir/9-3/paper181.html.

International Organization for Standardization. (1999). *ISO 13407: Human-centred design processes for interactive systems*.

Irani, Z., & Love, P. E. D. (2001). The propagation of technology management taxonomies for evaluating investments in information systems. *Journal of management information systems, 17*(3), 161-177.

Irani, Z., Love, P. E. D., Elliman, T., Jones, S., & Themistocleous, M. (2005). Evaluating e-government: learning from the experiences of two UK local authorities. *Information systems Journal, 15*(1), 61-82.

Jameson, A. (2003). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.), *Human-computer interaction handbook* (pp. 305-330). Mahwah: Erlbaum.

Jameson, A. (2007). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.), *Human-computer interaction handbook* (pp. 433-458). Mahwah: Erlbaum.

Jameson, A. (2009). Understanding and dealing with usability side effects of intelligent processing. *AI magazine, 30*(4), 23-41.

Jameson, A., & Schwarzkopf, E. (2002). Pros and cons of controllability: an empirical study. In P. De Bra, P. Brusilovsky & R. Conejo (Eds.), *Adaptive Hypermedia 2002* (pp. 193-202). Heidelberg: Springer.

Jaspers, M. W. M. (2009). A comparison of usability methods for testing interactive health technologies: Methodological aspects and empirical evidence. *International journal of medical informatics, 78*(5), 340-353.

Jeffries, R., Miller, J. R., Wharton, C., & Uyeda, K. M. (1991). *User interface evaluation in the real world: A comparison of four techniques*. Paper presented at the SIGCHI conference on Human factors in computing systems, April 27 - May 2, New Orleans, LA, USA.

Jensen, A. L., Boll, P. S., Thysen, I., & Pathak, B. K. (2000). Pl@nteInfo: A web-based system for personalised decision support in crop management. *Computers and electronics in agriculture, 25*(3), 271-293.

Jokela, T. (2001). *Assessment of user-centred design processes as a basis for improvement action*. University of Oulu, Oulu, Finland.

Kaasinen, E. (2003). User needs for location aware mobile services. *Personal ubiquitous computing, 7*(1), 70-79.

Kara, A., & Kaynak, E. (1997). Markets of a single customer: Exploiting conceptual developments in market segmentation. *European journal of marketing, 31*(11/12), 873-895.

Karat, C. (1994). A business case approach to usability cost justification. In R. G. Bias & D. J. Mayhew (Eds.), *Cost-justifying usability* (pp. 45-70). New York: Morgan Kaufmann.

Karat, C., Brodie, C., Karat, J., Vergo, J., & Alpert, S. R. (2003). Personalizing the User Experience on ibm.com. *IBM systems journal, 42*(4), 686-701.

Katz, E., Blumler, J. G., & Gurevitch, M. (1973). Uses and gratifications research. *The public opinion quarterly, 27*(4), 509-523.

Kaufman, J. (2006). Practical usability testing. *Digital web magazine* Retrieved on September 7, 2009, from http://www.digital-web.com/articles/practical_usability_testing/

Kay, J. (2000). Stereotypes, student models and scrutability. In G. Gauthier, C. Frasson & K. VanLehn (Eds.), *Intelligent tutoring systems* (pp. 19-30). Berlin: Springer.

Kay, J. (2006). Scrutable adaptation: Because we can and must. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 11-20). Heidelberg: Springer.

Ketamo, H. (2003). Xtask: An adaptable learning environment. *Journal of computer assisted learning, 19*(3), 360-370.

Kim, Y. H., & Kim, D. J. (2005). *A study of online transaction self-efficacy, consumer trust, and uncertainty reduction in electronic commerce transaction*. Paper presented at the 38th Hawaii international conference on system sciences, January 3-6, Hawaii, USA.

Kinzie, M. B., Cohn, W. F., Julian, M. F., & Knaus, W. A. (2002). A user-centered model for web site design: needs assessment, user interface design, and rapid prototyping. *Journal of the American medical informatics association, 9*(4), 320-330.

Kitchenham, B. A., Pfleeger, S. L., Pickard, L. M., Jones, P. W., Hoaglin, D. C., El Emam, K., et al. (2002). Preliminary guidelines for empirical research in software engineering. *IEEE transactions on software engineering, 28*(8), 721-734.

Kjeldskov, J., Graham, C., Pedell, S., Vetere, F., Howard, S., Balbo, S., et al. (2005). Evaluating the usability of a mobile guide: The influence of location, participants and resources. *Behavior & information technology, 24*(1), 51-65.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International journal of human-computer studies, 60*(5/6), 599-620.

Klaassen, R., Karreman, J., & Van der Geest, T. (2006). Designing Government Portal Navigation Around Citizens' Needs. In M. A. Wimmer, H. J. Scholl, A. Grönlund & K. V. Andersen (Eds.), *EGOV 2006* (pp. 162-173). Heidelberg: Springer.

Knutov, E., De Bra, P., & Pechenizkiy, M. (2009). AH 12 years later: a comprehensive survey of adaptive hypermedia methods and techniques. *New review of hypermedia and multimedia, 15*(1), 5-38.

Kobsa, A. (2001). Generic user modeling systems. *User modeling and user-adapted interaction, 11*(1/2), 49-63.

Kobsa, A., Koenemann, J., & Pohl, W. (2001). Personalized hypermedia presentation techniques for improving online customer relationships. *The knowledge engineering review, 16*(2), 111-155.

Kolari, J., Laakko, T., Hiltunen, T., Ikonen, V., Kulju, M., Suihkonen, R., et al. (2004). *Context-aware services for mobile users (technology and user experiences)*. VTT technical research centre of Finland.

Kotler, P., & Armstrong, G. (1999). *Principles of marketing*. Upper Saddle River: Prentice-Hall.

Krenner, J. (2002). Reflections on the Requirements Gathering in an One-Stop Government Project. In R. Traunmüller & K. Lenk (Eds.), *EGOV 2002*. (pp. 124-128). Heidelberg: Springer.

Kujala, S. (2003). User involvement: a review of the benefits and challenges. *Behavior & information technology, 22*(1), 1-16.

Kushniruk, A. W., & Patel, V. L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of biomedical informatics, 37*(56-76).

Labaw, P. J. (1981). *Advanced questionnaire design*. Cambridge: Abt books.

Lauesen, S. (2002). *Software Requirements. Styles and Techniques*. London: Addison-Wesley.

Lee, J., & Allaway, A. (2002). Effects of personal control on adoption of self-service technology innovations. *Journal of services marketing, 16*(6), 553-572.

Lefebvre, R. C., & Flora, F. J. (1988). Social marketing and public health intervention. *Health education quarterly, 15*(3), 299-315.

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

Lewis, R. (2005). Glossary of terms for device independence. Retrieved on May 6, 2010, from http://www.w3.org/TR/di-gloss/

Lines, L., Ikechi, O., & Hone, K. S. (2007). Accessing e-Government Services: Design Requirements for the Older User. In C. Stephanidis (Ed.), *Universal Access in HCI* (pp. 932-940). Heidelberg: Springer.

Liu, Y. (2003). Developing a scale to measure the interactivity of websites. *Journal of advertising research, 43*(2), 207-216.

Liu, Y., & Shrum, L. J. (2002). What is interactivity and is it always such a good thing? Implications of definition, person, and situation for the influence of interactivity on advertising effectiveness. *Journal of advertising, 31*(4), 53-64.

Livingstone, S. (2003). The changing nature of audiences: From the mass audience to the interactive media user. In A. N. Valdivia (Ed.), *A companion to media studies* (pp. 337-359). Malden: Blackwell publishing.

Lusoli, W., & Miltgen, C. (2009). *Young people and emerging digital services. An exploratory survey on motivations, perceptions and acceptance of risks*. Luxemburg: Office for official publications of the European communities.

Magoulas, G., Chen, S., & Papanikolaou, K. (2003). *Integrating layered and heuristic evaluation for adaptive learning environments*. Paper presented at the 2nd workshop on empirical evaluation of adaptive systems, held in conjunction with the 9th international conference of user modeling, June 22, Pittsburgh, PA, USA.

Maguire, M. (2001). Methods to support human-centred design. *International journal of human-computer studies, 55*(4), 587-634.

Malhotra, N. K., Kim, S. S., & Agarwal, J. (2004). Internet users' information privacy concerns (IUIPC): the construct, the scale, and a causal model. *Information systems research, 15*(4), 336-355.

Mao, J., Vredenburg, K., Smith, P. W., & Carey, T. (2005). The state of user-centered design practice. *Communications of the ACM, 48*(3), 105-109.

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *The academy of management review, 20*(3), 709-734.

McGraw, K. L., & Harbison, K. (1997). *User-centered requirements: the scenario-based engineering process*. Mahwah: Lawrence Erlbaum Associates.

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). The impact of initial consumer trust on intentions to transact with a web site: a trust building model. *Journal of strategic information systems, 11*(3/4), 297-323.

McQuail, D. (1997). *Audience analysis*. Thousand Oaks: Sage.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks: Sage.

Moon, J. W., & Kim, Y. G. (2001). Extending the TAM for a World-Wide-Web context. *Information & management, 38*(4), 217-230.

Muntean, C. H., & McManis, J. (2006). The value of QoE-based adaptation approach in educational hypermedia: empirical evaluation. In V. Wade, H. Ashman & B. Smyth (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 121-130). Berlin: Springer.

Nahl, D. (1998). Ethnography of novices' first use of web search engines: Affective control in cognitive processing. *Internet reference services quarterly, 3*(2), 51-72.

Napoli, P. M. (2008). *Toward a model of audience evolution: New technologies and the transformation of media audiences*. Bronx: The Donal McGannon communication research center.

NHS Centre for reviews and dissemination. (2001). *Undertaking systematic reviews of research on effectiveness*. York: University of York.

Niu, W. T., & Kay, J. (2008). Pervasive personalisation of location information: personalised context ontology. In W. Nejdl, J. Kay, P. Pu & E. Herder (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 143-152). Berlin: Springer.

Norman, D. A. (1986). *The design of everyday things*. New York: Basic books.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Olivero, N., & Lunt, P. (2004). Privacy versus willingness to disclose in e-commerce exchanges: The effect of risk awareness on the relative role of trust and control. *Journal of economic psychology, 25*(2), 243-262.

Oostveen, A., & Van de Besselaar, P. (2004). *From small scale to large scale user participation: a case study of participatory design in e-government systems*. Paper presented at the 8th conference on participatory design, July 27-31, Toronto, Canada.

Organisation for Economic Co-operation and Development. (2007). *Participative web and user-created content*. Paris: Organisation for Economic Co-operation and Development.

Pahnila, S. (2006). *Assessing the usage of personalized web information systems*. University of Oulu, Oulu, Finland.

Paramythis, A. (2009). *Adaptive systems: Development, evaluation and evolution*. Johannes Kepler University, Linz, Austria.

Paramythis, A., & Weibelzahl, S. (2005). A decomposition model for the layered evaluation of interactive adaptive systems. In L. Ardissono, P. Brna & A. Mitrovic (Eds.), *Proceedings of the 10th international conference on user modeling* (pp. 438-442). Heidelberg: Springer.

Paramythis, A., Weibelzahl, S., & Masthoff, J. (2010). Layered evaluation of interactive adaptive systems: framework and formative methods. *User modeling and user-adapted interaction, 20*(5), 383-453.

Pateli, A. G., Giaglis, G. M., & Spinellis, D. D. (2005). Trial evaluation of wireless info-communication and indoor location-based services in ex-

hibition shows. In P. Bozanis & E. N. Houstis (Eds.), *Advances in informatics* (pp. 199-210). Berlin: Springer.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks: Sage.

Peleg, M., Shackak, A., Wang, D., & Karnieli, E. (2009). Using multi perspective methodologies to study users' interactions with the prototype front end of a guideline-based decision support system for diabetic foot care. *International journal of medical informatics, 78*(7), 482-493.

Peters, J. D. (1999). *Speaking into the air. A history of the idea of communication*. Chicago: University of Chicago Press.

Petrelli, D. (2008). On the role of user-centred evaluation in the advancement of interactive information retrieval. *Information processing and management, 44*(1), 22-38.

Phelps, J., Nowak, G., & Ferrell, E. (2000). Privacy concerns and consumer willingness to provide personal information. *Journal of public policy & marketing, 19*(1), 27-41.

Pianesi, F., Graziola, I., & Zancanaro, M. (2007). Intrinsic motivational factors for the intention to use adaptive technology: Validation of a causal model. In C. Conati, K. McCoy & G. Paliouras (Eds.), *User modeling 2007* (pp. 258-267). Berlin: Springer.

Pianesi, F., Graziola, I., Zancanaro, M., & Goren-Bar, D. (2009). The motivational and control structure underlying the acceptance of adaptive museum guides - An empirical study. *Interacting with computers, 21*(3), 186-200.

Pieterson, W., Ebbers, W., & Van Dijk, J. (2007). Personalization in the public sector. An inventory of organizational and user obstacles towards personalization of electronic services in the public sector. *Government information quarterly, 24*(1), 148-164.

Plato. (trans. 2005). *Phaedrus* (C. Rowe, Trans.). London: Penguin books.

Robertson, S., & Robertson, J. (2006). *Mastering the Requirements Process* (2nd ed.). New York: Addison-Wesley.

Rothensee, M. (2008). User acceptance of the intelligent fridge: empirical results from a simulation. In C. Floerkemeier, M. Langheinrich, E. Fleisch, F. Mattern & S. E. Sarma (Eds.), *The Internet of things* (Vol. LNCS 4952, pp. 123-139). Berlin: Springer.

Rudd, J., Stern, K., & Isensee, S. (1996). Low vs. high-fidelity prototyping debate. *Interactions, 3*(1), 76-85.

Saiedian, H., & Dale, R. (2000). Requirements engineering: making the connection between the software developer and customer. *Information and Software Technology, 42*(6), 419-428.

Sandberg, K. W., & Pan, Y. (2007). The role of human factors in design and implementation of electronic public information systems. In D. Harris (Ed.), *Engineering psychology and cognitive ergonomics, HCII 2007* (pp. 164-173). Berlin: Springer.

Savage, P. (1996). *User interface evaluation in an interative design process: A comparison of three techniques*. Paper presented at the ACM conference on human factors, April 13-18, Vancouver, Canada.

Schmidt-Belz, B., & Posland, S. (2003). *User validation of a mobile tourism services*. Paper presented at the Workshop on HCI mobile guides, held in conjunction with MobileHCI03, September 8-11, Udine, Italy.

Schoenbachler, D. D., & Gordon, G. L. (2002). Trust and customer willingness to provide information in database-driven relationship marketing. *Journal of interactive marketing, 16*(3), 2-16.

Schwendtner, C., König, F., & Paramythis, A. (2006). *Prospector: An adaptive front-end to the Google search engine*. Paper presented at the 14th workshop on adaptivity and user modeling in interactive systems, held in conjunction with LWA, October 9-11, Hildesheim, Germany.

Scott, D. B. (2008). Assessing text processing: A comparison of four methods. *Journal literacy research, 40*(3), 290-316.

Smith, H., Fitzpatrick, G., & Rogers, Y. (2004). *Eliciting reactive and reflective feedback for a social communication tool: A multi-session approach*. Paper presented at the 5th conference on Designing interactive systems: processes, practices, methods, and techniques August 1-4, Cambridge, MA, USA.

Smith, W. R. (1956). Product differentation and market segmentation as alternative marketing strategies. *Journal of marketing, 21*(1), 3-8.

Snyder, C. (2003). *Paper prototyping*. San Francisco: Morgan Kaufmann.

Södergard, C., Aaltonen, M., Hagman, S., Hiirsalmi, M., Järvinen, T., Kaasinen, E., et al. (1999). Integrated multimedia publishing: combining TV and newspaper content on personal channels. *Computer networks, 31*(11-16), 1111-1128.

Soufi, B., & Maguire, M. (2007). Achieving Usability within e-government web sites illustrated by a case study evaluation. In M. J. Smith & G. Salvendy (Eds.), *Human interface, part II, HCII 2007* (pp. 777-784). Berlin: Springer.

Spector, P. E. (1992). *Summated rating scale construction*. Thousand Oaks: Sage.

Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1997). *Web site usability: A designer's guide*. North Andover: User interface engineering.

Stary, C., & Totter, A. (2003). Measuring the adaptability of universal accessible systems. *Behavior & information technology, 22*(2), 101-116.

Statistics Netherlands (2010). Key figures. Retrieved on November 9, 2010, from http://www.cbs.nl/en-GB/menu/cijfers/kerncijfers/default.htm?Languageswitch=on

Stein, A. (1997). Usability and assessments of multimodal interaction in the SPEAK! system: An experimental case study. *The new review of hypermedia and multimedia, 3*(1), 159-180.

Stock, O., Zancanaro, M., Busetta, P., Callaway, C., Krüger, A., Kruppa, M., et al. (2007). Adaptive, intelligent presentation of information for the museum visitor in PEACH. *User modeling and user-adapted interaction, 17*(3), 257-304.

Strategic Business Insights (2009). U.S. framework and VALS types. Retrieved on February 17, 2010, from http://www.strategicbusinessinsights.com/vals/ustypes.shtml

Sutcliffe, A. (1996). A conceptual framework for requirements engineering. *Requirements engineering, 1*(3), 170-189.

Sutcliffe, A. (1997). *A technique combination approach to requirements engineering*. Paper presented at the 3rd IEEE international symposium on requirements engineering, April 6-10, Annapolis, MD, USA.

Tam, K. Y., & Ho, S. Y. (2006). Understanding the impact of web personalization on user information processing and decision outcomes. *MIS Quarterly, 30*(4), 865-890.

Tauder, A. R. (2005). Getting ready for the next generation of marketing communications. *Journal of advertising research, 45*(1), 5-8.

Tintarev, N., & Masthoff, J. (2008). The effectiveness of personalized movie explanations: An experiment using commercial meta-data. In W. Nejdl, J. Kay, P. Pu & E. Herder (Eds.), *Adaptive hypermedia and adaptive web-based systems* (pp. 204-213). Berlin: Springer.

Tsakonas, G., & Papatheodorou, C. (2006). Analysing and evaluating usefulness and usability in electronic information services. *Journal of information science, 32*(5), 400-419.

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

Van der Geest, T. (2004). Beyond accessibility: Comparing three website usability test methods for people with impairments. In A. Dearden & L. Watts (Eds.), *Proceedings of HCI 2004: Design for life* (pp. 129-132).

Van der Geest, T., Jansen, J., Mogulkoç, E., De Vries, P., & De Vries, S. (2008). *Segmentation and e-government: A literature review*. Enschede: Telematica institute.

Van Oostendorp, H., & De Mul, S. (1999). Learning by exploration: Thinking-aloud while exploring an information system. *Instructional science, 27*(3/4), 269-284.

Van Velsen, L., Huijs, C., & Van der Geest, T. (2008). Eliciting user input for requirements on personalization: The case of a Dutch ERP system. *International journal of enterprise information system, 4*(4), 34-46.

Van Velsen, L., König, F., & Paramythis, A. (2009). *Assessing the effectiveness and usability of personalized internet search through a longitudinal evaluation*. Paper presented at the 6th workshop on user-centred design and evaluation of adaptive systems, held in conjunction with the international conference on user modeling, adaptation and personalization, June 26, Trento, Italy.

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

Virzi, R. A., Sokolov, J. L., & Karis, D. (1996). *Usability problem identification using both low- and high-fidelity prototypes*. Paper presented at the SIGHI conference on human factors in computing systems: Common ground, April 13-18, Vancouver, Canada.

Vredenburg, K., Mao, J., Smith, P. W., & Carey, T. (2002). *A survey of user-centered design practice*. Paper presented at the SIGCHI conference on human factors in computing systems, April 20-25, Minneapolis, MN, USA.

Walker, M., Takayama, L., & Landay, J. A. (2002). *High-fidelity or low-fidelity, paper or computer? Choosing attributes when testing web prototypes*. Paper presented at the 46th annual meeting of the human factors and ergonomics society, September 30 - October 4, Baltimore, MD, USA.

Wang, L., Bretschneider, S., & Gant, J. (2005). *Evaluating web-based e-government services with a citizen-centric approach*. Paper presented at the 38th Hawaii international conference on system sciences, January 3-6, Hawaii, USA.

Webster, J. G. (1998). The audience. *Journal of broadcasting & electronic media, 42*(2), 190-207.

Weibelzahl, S. (2003). *Evaluation of adaptive systems*. Pedagogical university Freiburg, Freiburg, Germany.

Weibelzahl, S. (2005). Problems and pitfalls in the evaluation of adaptive systems. In S. Y. Chen & G. D. Magoulas (Eds.), *Adaptable and adaptive hypermedia systems* (pp. 285-299). Hershey: IRM Press.

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

Welle Donker-Kuijer, M., De Jong, M., & Lentz, L. (2008). Heuristic web site evaluation. *Technical communication, 55*(4), 392-404.

Wharton, C., Rieman, J., Lewis, C., & Polson, P. (1994). The cognitive walkthrough method: a practitioner's guide. In J. Nielsen & R. L. Mack (Eds.), *Usability inspection methods* (pp. 105-140). New York: John Wiley & sons, Inc.

Wimmer, M. A., & Holler, U. (2003). Applying a Holistic Approach to Develop User-Friendly, Customer-Oriented e-Government Portal Interfaces. In N. Carbonell & C. Stephanidis (Eds.), *User Interfaces for All.* (pp. 167-178). Heidelberg: Springer.

Wu, D., Im, I., Tremaine, M., Instone, K., & Turoff, M. (2003). *A framework ofr classifying personalization scheme used on e-commerce websites*. Paper presented at the 36th Hawaii international conference on system sciences, January 6-9, Hawaii, USA.

Zabed Ahmed, S. M. (2008). A comparison of usability techniques for evaluating information retrieval system interfaces. *Performance measurement and metrics, 9*(1), 48-58.

Zhang, D. (2007). Web content adaptation for mobile handheld devices. *Communications of the ACM, 50*(2), 75-79.

Zhang, P., & Von Dran, G. M. (2001). User expectations and rankings of quality factors in different web site domains. *International journal of electronic commerce, 6*(2), 9-33.

Zimmermann, A., & Lorenz, A. (2008). LISTEN: A user-adaptive audio-augmented museum guide. *User modeling and user-adapted interaction, 18*(5), 389-416.

# Samenvatting

**(Summary in Dutch)**

Steeds meer organisaties bieden hun klanten of cliënten op maat gemaakte digitale communicatie aan. Deze techniek, ook wel personalisering genoemd, is bij het grote publiek bekend in de vorm van Amazon boekaanbevelingen of de persoonlijke startpagina *i*Google. Bij personalisering wordt digitale communicatie op het individu en haar karakteristieken, voorkeuren en context afgestemd. Deze afstemming wordt gebaseerd op een gebruikersmodel: een bestand waarin deze kenmerken zijn opgeslagen. Personalisering kan vele vormen aannemen. Zo kan de inhoud van een boodschap worden gepersonaliseerd (door fragmenten toe te voegen of te verwijderen), kan de presentatie worden aangepast (bijvoorbeeld door geen plaatjes te tonen als een website wordt bekeken op een mobiele telefoon), enzovoort.

Als een organisatie besluit om gebruik te maken van personalisering in haar digitale communicatie zal het ontwerpteam rekening moeten houden met een aantal zaken tijdens het gehele ontwerptraject. Zo zijn er een aantal onderdelen van de gebruikerservaring waar extra aandacht aan moet worden geschonken. Hebben ontvangers nog wel het idee dat ze controle hebben over de selectie van informatie op een gepersonaliseerde website? Wordt er geen inbreuk op de privacy gepleegd door persoonlijke informatie op te slaan? En wordt een klant niet de kans ontnomen om leuke, nog onbekende boeken te ontdekken als aanbevelingen gebaseerd zijn op mijn koopgedrag uit het verleden? Daarnaast moet uit evaluaties blijken of gepersonaliseerde communicatie goed is afgestemd op het individu. Maar hoe doe je dit als iedereen een andere boodschap te zien krijgt?

Gebruikersgericht ontwerpen kan een zeer geschikte ontwerpaanpak zijn voor personalisering. De filosofie achter deze ontwerpaanpak is dat vanaf het begin van het ontwerpproces de toekomstige gebruiker en haar taken centraal moeten staan, dat constant evaluaties van (tussenversies) van de digitale communicatie moeten worden uitgevoerd, en dat iteratief ontwerp moet worden toegepast. In dit proefschrift presenteer ik vier studies die bijdragen aan het arsenaal van ontwerp- en evaluatiemethoden van ontwerpteams die een gebruikersgerichte ontwerpaanpak hanteren voor het ontwerpen van gepersonaliseerde digitale communicatie. Daarnaast presenteer ik enkele concrete ontwerprichtlijnen.

In hoofdstuk 2 bespreek ik een studie die de rol van vertrouwen en het gevoel van controle verkent in de totstandkoming van de beslissing om wel of geen gebruik te maken van een gepersonaliseerd aanbod van informatie op websites. Een dergelijk onderzoek hoort plaats te vinden in een zeer vroege fase in het ontwerpproces: de verkenning van de gebruikerscontext. In een online experiment werden 1.141 deelnemers een standaard, niet gepersonaliseerde pagina van een gemeentelijke website getoond, waarna vier

van de vijf deelnemers ook één van vier gepersonaliseerde varianten te zien kreeg. Aan het begin van het online experiment werd het vertrouwen in de gemeente vastgesteld via een vragenlijst. Aan het einde werd, wederom door middel van een online vragenlijst, voor elke variant het vertrouwen in de technologie, het gevoel van controle, en de intentie om een dergelijke techniek te gebruiken vastgesteld. Vertrouwen in de organisatie had geen invloed op de gebruiksintentie, terwijl vertrouwen in de technologie voor elke variant van personalisering wel van invloed was. Het gevoel van controle bleek een zeer belangrijke invloed op deze beslissing te hebben. Dit gevoel van controle was het grootst voor de vorm van personalisering waarbij de gebruiker zelf mag kiezen welke informatie getoond wordt (adaptability; zoals in *i*Google). Bij het ontwerpen van een gepersonaliseerd informatie aanbod op websites moet men dus vooral functionaliteiten toevoegen die de gebruiker in staat stelt om de regie te voeren over het selectieproces.

In hoofdstuk 3 presenteer ik een gebruikersgerichte aanpak voor het identificeren en formuleren van ontwerpeisen voor gepersonaliseerde elektronische dienstverlening van de overheid. Dit zijn activiteiten die plaatsvinden in een volgende fase in het gebruikersgerichte ontwerpproces: het vaststellen van ontwerpeisen. De gepresenteerde aanpak maakt gebruik van interviews, een methode om ontwerpeisen te formuleren die een extra nadruk legt op meetbare succescriteria, het ontwerpen van ruwe prototypes en evaluaties door middel van demonstraties en interviews met burgers. Deze aanpak is gevalideerd door middel van een case study. Deze bevestigde het belang van iteratief design, aangezien de vertaling van uitspraken van toekomstige gebruikers in ontwerpeisen en vervolgens in ontwerp niet altijd strookt met de originele karakteristieken, voorkeuren en contexten van gebruikers. De case study toonde tot slot het belang van een multidisciplinair ontwerpteam voor het ontwerpen van gepersonaliseerde, elektronische dienstverlening.

Na het ontwerp van een (tussenversie van) gepersonaliseerde digitale communicatie is het zaak om verbeterpunten op te sporen (formatieve evaluatie) of het effect van de communicatie vast te stellen (summatieve evaluatie). Dit zijn de laatste stappen uit het gebruikersgerichte ontwerpproces. In hoofdstuk 4 staat beschreven hoe een literatuur review is uitgevoerd naar gebruikersgerichte evaluatie van personalisering. Dit zijn evaluaties waarin gebruikerservaringen van personalisering worden vastgesteld of problemen met betrekking tot gebruikersgemak worden opgespoord. De resultaten tonen aan dat evaluaties die worden gerapporteerd in de wetenschappelijke literatuur niet in lijn zijn met de filosofie achter gebruikersgericht ontwerpen. Vragenlijsten zijn zeer populair, terwijl probleemopsporende methoden

(zoals hardop denken) nauwelijks worden toegepast. In de laatste jaren is een toename te zien van studies die zich richten op acceptatie van personalisering en vertrouwen, of iteratief ontwerp bespreken. Deze trend suggereert dat gebruikersgericht ontwerpen van personalisering aan populariteit wint in de wetenschappelijke gemeenschap.

In hoofdstuk 5 vergelijk ik het nut van drie methoden (interviews, vragenlijsten met open vragen en hardop denken) voor de formatieve evaluatie van personalisering. Uit deze vergelijking blijkt dat hardop denken de enige methode is die alle kritieke en serieuze problemen die gerelateerd zijn aan personalisering blootlegt. Daarnaast is dit ook de methode die het beste commentaar op de gepercipieerde kwaliteit van personalisering ontlokt bij proefpersonen. De mening van deelnemers over de specifieke onderdelen van de gebruikerservaring die extra aandacht vereisen bij gepersonaliseerde digitale communicatie (zoals privacy, een gevoel van controle, enzovoort) blijkt het best ontlokt te kunnen worden met behulp van vragenlijsten met open vragen. Ik concludeer dat bij een formatieve evaluatie van gepersonaliseerde digitale communicatie het best kan worden gekozen voor hardop denken, aangevuld met vragenlijsten met open vragen die specifiek ingaan op onderdelen van de gebruikerservaring die van belang zijn bij gepersonaliseerde digitale communicatie.

In het laatste hoofdstuk van dit proefschrift bespreek ik ten eerste een technische ontwerpaanpak van personalisering: 'layered evaluation'. Deze aanpak heeft als voornaamste doel om het proces van personalisering (verzameling van data, interpretatie van data, keuze van geschikte digitale communicatie voor individu) te optimaliseren door middel van ontwerpstudies en evaluaties tijdens het gehele ontwerpproces. Layered evaluation en gebruikersgericht ontwerpen hebben een eigen insteek, maar overlappen elkaar ook deels. In de toekomst moeten beide aanpakken geïntegreerd worden zodat gepersonaliseerde digitale communicatie kan worden ontworpen die technisch optimaal is en voldoet aan de wensen en eisen van gebruikers.

Uit verschillende publicaties komt de opvatting naar voren dat personalisering van digitale communicatie altijd beter is dan geen personalisering, en dat een impliciete vorm (adaptivity) beter is dan een expliciete (adaptability). Naar aanleiding van verschillende studies waarin de visie van gebruikers over het nut van personalisering wordt vastgesteld blijkt dat *een bepaalde vorm* van personalisering beter *kan* zijn dan geen personalisering, *afhankelijk van de taak die de gebruiker uit wil voeren*. Het is daarom van groot belang dat wordt onderzocht bij welke soort systeem en welke soort taak personalisering een toegevoegde waarde kan hebben.

Tot slot stip ik aan dat het huidige wetenschappelijke onderzoek zich voornamelijk focust op effectiviteit en efficiëntie van personalisering, terwijl de gebruikerservaring achterwege blijft. Ik beargumenteer dat hoge effectiviteit en efficiëntie niet automatisch leiden tot gebruiksvriendelijke en prettige gepersonaliseerde digitale communicatie. Daarom moet onderzoek zich meer richten op de gebruiksvriendelijkheid en gebruikerservaring van personalisering.

# Bibliography

## Journal articles

Van Velsen, L., Van der Geest, T., Van de Wijngaert, L., Van den Berg, S.M. & Steehouder, M. The role of trust and controllability in user acceptance of online content personalization. *In review.*

Van Velsen, L., Van der Geest, T. & Klaasen, R. Identifying usability issues for personalization during formative evaluations: A comparison of three methods. *To appear in the International Journal of Human-Computer Interaction.*

Melenhorst, M. & Van Velsen, L. Tempting to Tag: An Experimental Comparison of Four Tagging Input Mechanisms. *Human Technology*, *6*(2), 229-248.

Van Velsen, L. Van der Geest, T. & Steehouder, M. (2010). The contribution of Technical Communicators to the User-Centered Design Process of Personalized Systems. *Technical Communication, 57*(2), 182-196.

Van Velsen, L. & Melenhorst, M. (2009). Incorporating user motivations to design for video tagging. *Interacting with Computers, 21*(3), 221-232.

Van Velsen, L., Van der Geest, T., Ter Hedde, M. & Derks, W. (2009). Requirements engineering for e-Government services: A citizen-centric approach and case study. *Government Information Quarterly, 26*(3), 477-486.

Van Velsen, L., Huijs, C. & Van der Geest, T. (2008). Eliciting User Input for Requirements on Personalization: The Case of a Dutch ERP System. *International Journal of Enterprise Information Systems, 4*(4), 34-46.

Van Velsen, L., Van der Geest, T., Klaassen, R. & Steehouder, M. (2008). User-centered evaluation of adaptive and adaptable systems: a literature review. *The Knowledge Engineering Review, 23*(3), 261-281.

Van Velsen, L.S., Steehouder, M.F. & De Jong, M.D.T. (2007). Evaluation of User Support. Factors that affect User Satisfaction with Helpdesks and Helplines. *IEEE Transactions on Professional Communication, 50*(3), 219-213.

## Book chapters

Van Velsen, L., Huijs, C. & Van der Geest, T. (2010). Requirements elicitation for personalized ERP systems: A case study. In A. Gunasekaran & T. Shea (Eds.), *Organizational advancements through enterprise information systems: Emerging applications and developments* (pp. 46-56). Hershey, PA: IGI Global.

## Conference papers

König, F., Van Velsen, L. & Paramythis, A. (2009). Finding My Needle in the Haystack: Effective Personalized Re-ranking of Search Results in Prospector. In T. Di Noia & F. Buccafurri (Eds.), *E-Commerce and Web Technologies 2009, LNCS 5692* (pp. 312-323). Heidelberg: Springer.

Van Velsen, L., Van der Geest, T., Ter Hedde, M. & Derks, W. (2008). Engineering User Requirements for e-Government Services: A Dutch Case Study. In M.A. Wimmer, H.J. Scholl & E. Ferro (Eds.), *Electronic Government 2008, LNCS 5184* (pp. 243-254). Heidelberg: Springer.

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

## Editorship

T. Van der Geest, and L. van Velsen (Eds.) Proceedings of the IEEE Professional Communication Conference, held in Enschede, the Netherlands, July 7-9, 2010.

S. Weibelzahl, J. Masthoff, A. Paramythis, and L. van Velsen (Eds.) Proceedings of the Sixth Workshop on User-Centred Design and Evaluation of Adaptive Systems, held in conjunction with the International Conference on User Modeling, Adaptation, and Personalization (UMAP 2009), Trento, Italy, June 26th, 2009.

## Workshop papers

Hauger, D. & Van Velsen, L. (2009). *Analyzing client-side interactions to determine reading behavior*. Paper presented at the 17th Workshop on adaptivity and user modeling in interactive systems (ABIS), Darmstadt, Germany.

Van Velsen, L., König, F. & Paramythis, A. (2009). *Assessing the effectiveness and usability of personalized internet search through a longitudinal evaluation*. Paper presented at the Sixth workshop on user-centred design and evaluation of adaptive systems, Trento, Italy.

Paramythis, A., König, F., Schwendtner, C. & Van Velsen, L. (2008). *Using thematic ontologies for user– and group-based adaptive personalization in web searching*. Paper presented at Adaptive Multimedia Retrieval, Berlin, Germany.

Van Velsen, L. & Melenhorst, M. (2008). *User Motives for Tagging Video Content*. Paper presented at the Adaptation for the Social Web workshop, Hannover, Germany.

Van Velsen, L.S., Van der Geest, T.M. & Klaassen, R.F. (2006). *User-Centered Evaluation of Adaptive and Adaptable Systems*. Paper presented at the fifth workshop on User-Centred Design and Evaluation of Adaptive Systems. June 20th, Dublin, Ireland.

## Poster presentations

Van Velsen, L., Van der Geest, T & Ter Hedde, M (2007). *Supporting e-Service Clients with Personalized Narratives.* Poster presented at the IEEE Professional Communication Conference, Seattle, USA.

# Dankwoord

## (Acknowledgement in Dutch)

Op de voorkant van een proefschrift staat altijd maar één naam. Als promovendus mag je dan wel de meeste eer voor het werk opeisen, het resultaat was nooit tot stand gekomen zonder de hulp, adviezen, kritische opmerkingen en morele steun van een groep collega's, vrienden en familie. Ik wil hier dan ook graag de mensen bedanken zonder wie dit proefschrift er waarschijnlijk niet was gekomen.

Als eerste wil ik mijn dagelijks begeleider en co-promotor, Thea van der Geest, bedanken. Thea, ik heb de afgelopen jaren veel van je geleerd. Daarnaast heb je me de ruimte gegeven om mijn eigen projecten op te zetten en mijn eigen interesses achterna te gaan. Ik vond het erg leuk om de afgelopen jaren samen als 'advocaten van de gebruikers' aan zeer uiteenlopende projecten te werken. Dank hiervoor.

Michaël Steehouder wil ik graag bedanken voor zijn kritische en verfrissende blik op mijn tussenversies en ideeën. Ook heb je met je praktische aanpak de vaart erin gehouden. Misschien nog wel het belangrijkste is dat je me als student uit de collegebanken hebt gepikt om eerst je student-assistent te worden, daarna je afstudeerder en uiteindelijk je promovendus. Bedankt voor het bieden van deze kansen.

Ik wil de commissieleden bedanken voor het lezen van dit proefschrift en hun bijdrage aan de verdediging ervan. Daarnaast wil ik Peter de Vries en Wil Janssen bedanken voor hun waardevolle opmerkingen aan het begin van mijn promotie.

In de onderzoeken die ik heb uitgevoerd ben ik bijgestaan door veel mensen die verschillende, maar altijd belangrijke, bijdragen hebben geleverd. Hierbij wil ik mijn collega-onderzoekers Stéphanie van den Berg, Wijnand Derks, Marc ter Hedde, Rob Klaassen, Mark Melenhorst, Sanne ten Tije en Lidwien van de Wijngaert, bedanken voor hun inbreng. Speciale dank gaat uit naar Francien Malecki van Firmm, die voor de prachtige screenshots voor Hoofdstuk 2 zorgde. Tot slot wil ik de honderden deelnemers bedanken voor hun tijd en de kopjes koffie bij de interviews.

Mijn tijd als junior-onderzoeker en promovendus heb ik doorgebracht te midden van mijn collega´s bij de verschillende communicatiewetenschappen afdelingen. Dankzij hen ben ik elke dag met plezier naar mijn werk gegaan en had ik iemand om tegenaan te praten als het schrijven even niet wilde vlotten. In het bijzonder wil ik de volgende mensen bedanken: Alexander, Carolina, Joyce, Marieke, Mirjam, Nalini, Peter, Sanne, Thomas, Vanessa, Wendy en Willem. Emmy wil ik graag heel erg bedanken voor haar hulp tijdens de afronding van dit proefschrift. Ik vond het heel fijn dat je altijd voor me klaarstond als ik je hulp nodig had.

Ein besonderer Dank gilt meinen Kollegen an der Johannes Kepler Universität, Linz. Ich hatte eine phantastische Zeit bei Euch, mit Euch, den

Krapfen und den vielen Stunden Big Bang Theorie. Besondere Freude habe ich jedoch an der Tatsache daβ trotz der relativ großen Entfernung unsere Freundschaft standhielt. Danke Alexandros, Florian und Mirjam.

Mijn paranimfen Dimitri Geskus en Nicole Loorbach wil ik bedanken voor hun steun op 25 februari. Daarnaast wil ik Nicole heel erg bedanken voor de tijd die we samen als roomies hebben doorgebracht in Cubicus C145 en C228. Zonder je aanwezigheid en de goedkeurende blik van 'Big sister Linn' waren de afgelopen jaren lang niet zo leuk geweest.

Tot slot wil ik nog een aantal mensen bedanken die niet zo direct bij het onderzoek waren betrokken, maar minstens net zo belangrijk waren. Mijn moeder en broer wil ik bedanken voor de altijd welkome ontvangst in Nij-megen en hun steun op de momenten dat het allemaal wat minder lekker ging. Nora, jij hebt de ups en downs die horen bij het promoveren buiten werktijd mogen opvangen. Bedankt voor je steun, interesse, humor en lief-de.

Lex van Velsen has completed his PhD at the Department of Technical and Professional Communication at the University of Twente, the Netherlands.

His main research interests include personalization, user-centered design, e-service design and the Web 2.0.

This thesis discusses the implications of personalization for different phases of the user-centered design process: Analyzing the context of use, requirements engineering, design, and evaluation.