INTL. JOURNAL OF HUMAN–COMPUTER INTERACTION, *27*(7), 670–698, 2011

ISSN: 1044-7318 print / 1532-7590 online
DOI: 10.1080/10447318.2011.555304

Taylor & Francis
Taylor & Francis Group

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

Lex van Velsen, Thea van der Geest, and Rob Klaassen
University of Twente, Enschede, the Netherlands

A personalized system is one that generates unique output for each individual. As a result, personalization has transformed the interaction between the user and the system, and specific new usability issues have arisen. Methods used for evaluating personalized systems should be able to reveal the issues and problems specifically associated with personalization. Therefore this study evaluated three of the most common test methods used to detect usability problems in a personalized search engine. This was done by comparing the comments generated from thinking-aloud, questionnaires, and interviews. Questionnaires and interviews appear to be more useful for assessing specific usability issues for personalization, whereas thinking-aloud generates more comments on the usefulness of the system in the intended context of use and identifies the most critical and serious problems. Interviews, on the other hand, appear to yield a disproportionate number of positive comments. During the formative evaluation of a personalized system it is best to use a combination of thinking-aloud and questionnaires. This article concludes with a summary of implications for practitioners.

1. INTRODUCTION

In 1998, Jakob Nielsen described personalization as "much over-rated and mainly used as a poor excuse for not designing a navigable website" (para. 6). Since then,

We thank our colleagues from the Institute for Information Processing and Microprocessor Technology at the Johannes Kepler University in Linz, Austria, for making the Prospector system available for this study, as well as for their assistance and useful suggestions during the evaluation. We are also grateful to the external usability experts who helped in constructing interreliability scores.

Correspondence should be addressed to Lex van Velsen, Department of Psychology, Health and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, the Netherlands. E-mail: l.s.vanvelsen@utwente.nl

several websites have proven Nielsen wrong and have shown that personalization *can* be successful. For example,

- Amazon and Netflix have popularized personalized recommendations,
- Netvibes and *i*Google have enabled many Internet users to create their own personal home pages, and
- Museums have started to offer personalized tours through their collections.

Personalization is all about providing tailored output to individuals, based on their unique personal characteristics, needs, or contexts. Let us begin by explaining the concept of personalization in more detail. We can use the example of the personal recommendations made by the online publishing company Amazon to illustrate this. Imagine buying the first three Harry Potter books from Amazon. Amazon will automatically record these purchases in a file called *user profile*. A user profile is a collection of a user's personal characteristics, needs, and context. Amazon will then look for other customers with a similar user profile: people who also bought the first three Harry Potter books. Next, the system selects the items that most people similar to you already bought but which you yourself have not yet purchased. What completes the process of personalization is the communication of these selected items as "recommendations." Amazon can recommend, for example, the fourth Harry Potter book; J.K. Rowlings's new book, *The Tales of Beedle the Bard*; or the children's fantasy book *Eragon* by Christopher Paolini. Based on your purchasing behavior, these recommendations look logical and interesting. But what would a customer think if he or she was recommended the fourth to seventh Harry Potter books, all the Harry Potter books in hardcover, and the Harry Potter movies? Such recommendations would not help them to "discover something new" and are not very interesting. And what would happen if the system, on the basis of buying the first three Harry Potter books, recommends a book on weight loss? This would be rather surprising to say the least and might make the customer wonder whether Amazon's recommendations are useful at all. These examples of valueless or unpredictable recommendations make it clear that personalized systems need to comply with several usability guidelines that ensure successful personalization, besides the generic usability guidelines listed in numerous textbooks.

Jameson (2007) listed the specific usability issues for personalization:

- *Predictability*. Users must be able to predict the consequences of their actions for the generation of personalized output.
- *Comprehensibility*. Users must be able to understand how user profiling and the tailoring of system output works.
- *Controllability*. Users must be able to control their user profile and the generation of personalized output.
- *Unobtrusiveness*. Users must be able to complete their tasks without being distracted by personalization features.
- *Privacy*. Users must not have the feeling that the generation of a user profile infringes on their privacy.

- *Breadth of experience*. Users must not lose the possibility of discovering something new because output only complies with their user profile.
- *System competence*. Users must not have the feeling that the system creates an invalid user profile or does not personalize output successfully.

Evaluating these specific issues is particularly important during the formative evaluation process. This kind of evaluation focuses 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.

With the introduction of these new, specific usability issues, we have to ask ourselves two questions. Are the evaluation methods we have always used during formative evaluations suitable for eliciting user comments on these specific usability issues? Which method or combination of methods can we best use to identify problematic issues with the personalization provided to users? A related issue concerns the difficulty of assessing the user's perception of how "good" personalization is, and how it can be improved. Personalization is a process that often goes unnoticed by the user (Weibelzahl, 2005). This raises the question, "How can we elicit comments from formative evaluation participants about the quality of personalization?" So far, academic research has not produced any answers to these questions (Van Velsen, Van der Geest, Klaassen, & Steehouder, 2008). Yet, to design an effective formative evaluation, answers to these questions are crucial. The study presented in this article compares the suitability of the three main methods—thinking-aloud, questionnaires, and the interview—for eliciting comments first on the specific usability issues and second on user perceptions of the quality of personalization. With this information, evaluators can decide which method to apply in order to elicit comments that are necessary for creating usable personalization. It is the goal of this study to enable evaluators to design an effective formative evaluation of a personalized system.

2. A DISCUSSION ABOUT THREE DIFFERENT METHODS: THINKING-ALOUD, QUESTIONNAIRES, AND INTERVIEWS

Three methods appear to be popular in the scientific literature for evaluating personalization: thinking-aloud, questionnaires, and interviews (Van Velsen et al., 2008). We therefore focus our efforts on assessing their suitability for eliciting user comments on usability issues relating to personalization and the perceived quality of tailored output. Let us start by presenting a short overview of these methods.

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) claimed that, for personalized systems, thinking-aloud can

be conducted to elicit comments on users' cognitions and their thoughts on the usability of interface adaptations.

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 it is this type of information that is invaluable when one wants to improve the system. According to Labaw (1981), closed questions have another caveat: They do not give any indication of whether 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 (Henderson, Smith, Podd, & Varela-Alvarez, 1995; Miles & Huberman, 1994). They offer participants 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 (Carter, 2007). Gena and Weibelzahl (2007) claimed 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).

Interviews may be structured or semistructured. 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 semistructured 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 semistructured 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) claimed that, for the case of personalization, interviews are the most effective method for assessing user opinions and satisfaction levels.

Several studies have been published in which the value of different user-centered evaluation methods have been compared. The question that this body of literature might be able to answer is, What do we know about the value of thinking-aloud, interviews and questionnaires for eliciting comments on generic usability issues? The answer to this question can serve as input for our hypotheses.

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; Van der Geest, 2004). 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 managed to

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, whereas 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 advised 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. So how does this information answer the question we just posed? The literature suggests that for the case of generic usability issues, thinking-aloud seems to be the best method for identifying problems.

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 et al., 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. (1995), for example, found the added value of using other methods in combination with thinking-aloud to be limited. Therefore, during a comparison, it is important to assess the relative advantage of user-centered evaluation methods.

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 compare the comments on personalization elicited through thinking-aloud (a method applied during the actual process of interacting with the system), on 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

RQ: 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 and 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, whereas we would consider the "misunderstanding of the working of a drop-down menu" to be a generic issue.

We 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.

3.1. Specific Issues

The hypotheses we list for the methods' ability to elicit comments on specific issues are focused on the following dependent variables. Comments on

- the specific usability issues (which consist of the specific usability issues, listed by Jameson, 2007),
- the appreciation of personalization,
- the perceived quality of personalization (in this study seen as perceived relevance of search results), and
- the value of the comment (positive, negative or neutral).

These variables cover the range of comments evaluation participants can give that are related to personalization.

Our first hypothesis addresses the ability of three user-centered evaluation methods to elicit comments on specific usability issues and one related issue: appreciation of personalization. As no studies have delved into this matter before, our hypothesis 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. Because 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 et al., 2002). So, in accordance with a study by Ebling and John (2000), the set of issues identified by thinking-aloud, on 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.

3.2. Generic Issues

Of course, a formative evaluation of a personalized system needs to uncover a lot more than just specific issues. 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 (e.g., Lindgaard, 1994; Nielsen & Mack, 1994). As the goal of this study is not to compare the success of the three methods in eliciting comments on generic issues, we treat the results concerning generic issues generally. As a result, we did not formulate any hypotheses for these analyses.

4. METHOD

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

4.1. 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, the system is still in development so we were guaranteed that 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 explain next, 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. 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 metasearch engine that reranks 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 1), user ratings of search results are collected via a rating frame that is displayed above each opened search result (see Figure 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 to provide well-tailored output, relevant hits will appear higher in the list of search results than nonrelevant 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).

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, whereas 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

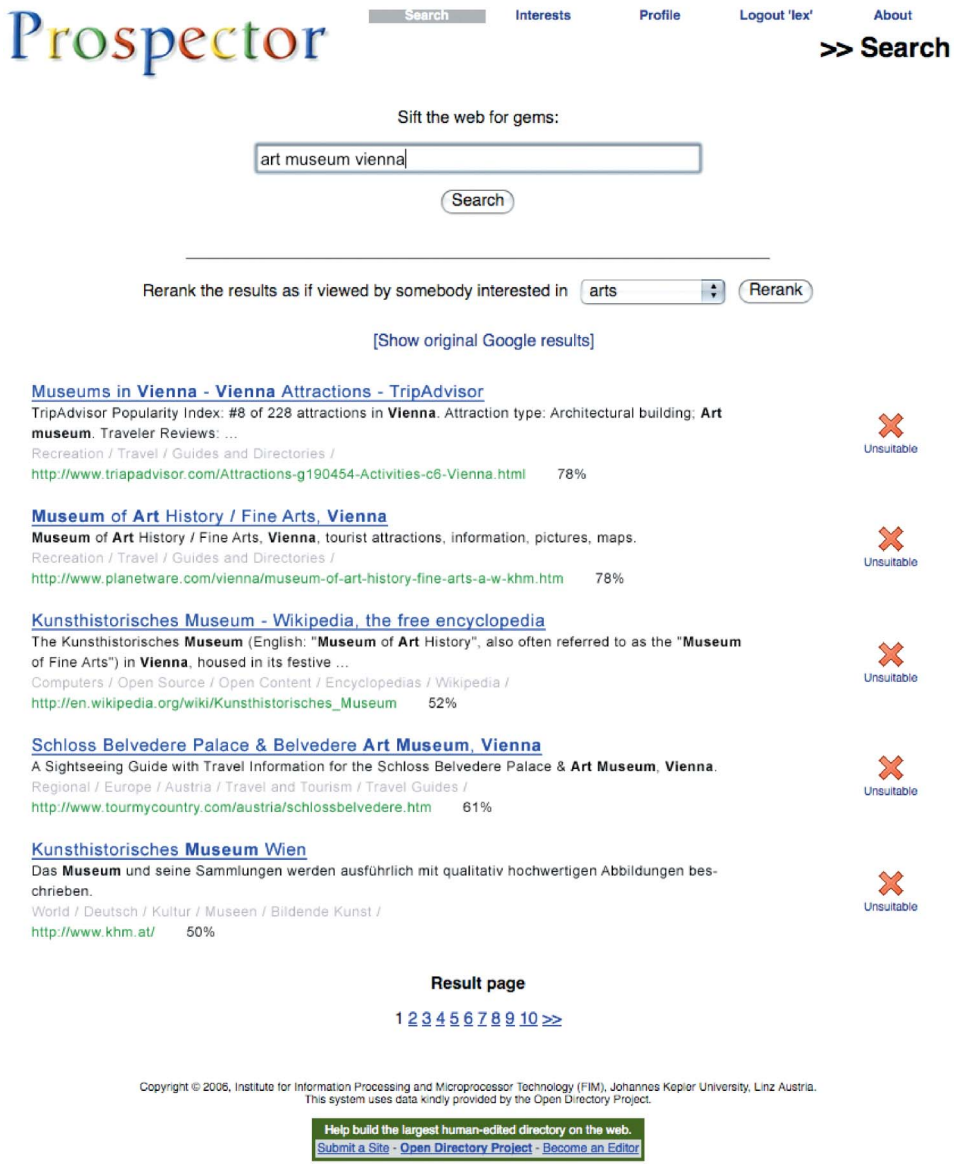

FIGURE 1 Prospector main screen (color figure available online).

FIGURE 2 Prospector rating frame (color figure available online).

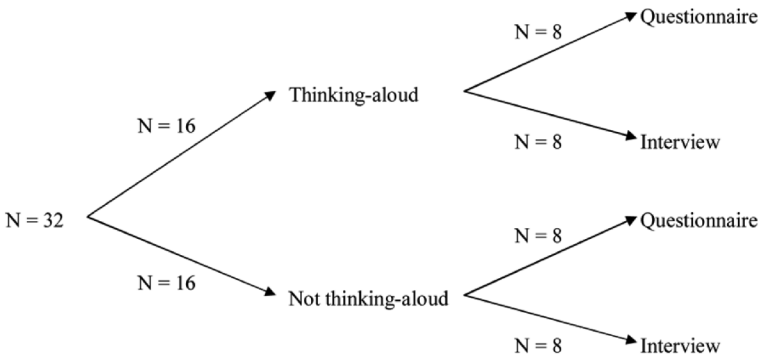

FIGURE 3 Conditions and participants.

questionnaire ($n = 8$). A pictorial overview of the participants and their conditions can be found in see Figure 3. This study design was chosen 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 is discussed in the Results section, no interaction effects occurred.

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 10 min per search. The Museum of Modern Art search was what Spool, Scanlon, Schroeder, Snyder, and DeAngelo (1997) called 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 were supplemented with data from observations, when necessary to clarify the audio recording.

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 nonpersonalized 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 because the interviews were semistructured, 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 4 displays the test procedure. The questionnaire and interview items can be found in the appendix.

4.3. Data Analysis

Comments on specific and generic issues were abstracted from the audio recordings and transcribed by one of the researchers. 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?

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

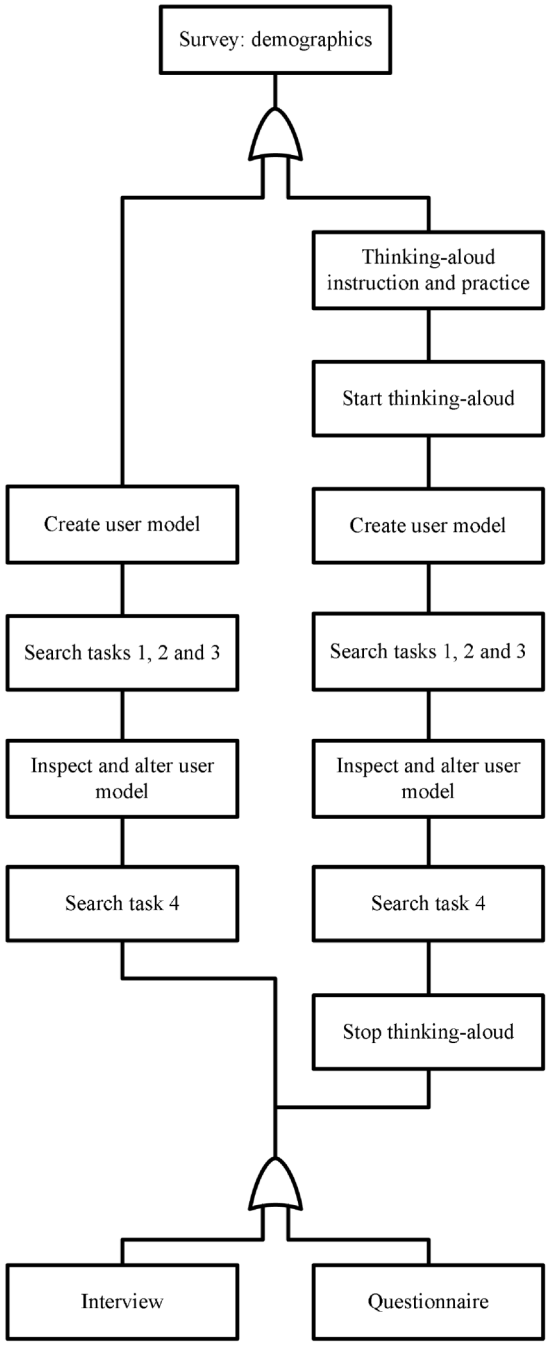

FIGURE 4 Study timeline.

problems are "similarity between Prospector and Google" or "understanding of keywords" in the user model.

We 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.

4.4. Pitfalls of Evaluating Personalized Systems

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 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, wishes, or context. When personalized output *does* provide a clear match, personalization might go unnoticed. This complication makes it clear that, during an evaluation, the perception of personalization should not be asked about directly (e.g., "Do you like personalization?"). It should be posed in terms of the variable it is supposed to serve. For example, when one is evaluating an online music store with personalized recommendations, one should not ask, "Do you like the personalized recommendations this music store gives you?" Rather, the question should be, "Is this recommended item one you would consider buying?" This last example refers to the perceived usefulness of a recommendation. 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 nonpersonalized 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 be evaluated fully only 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. RESULTS

The results of the analyses we performed after transcribing all the user feedback are described next. They enable us to answer our main research question: What is the yield of thinking-aloud, questionnaires or interviews when applied during the formative evaluation of a personalized system?

5.1. Participants

In total, 32 social science undergraduates participated in the evaluation. Twenty-four were female, and eight were male. They were 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 (nine 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 min 40 s to complete the test tasks. The other participants took on average 39 min 4 s.

5.2. Quality of Measurement

First we have to determine whether the coding of comments was correct. To determine the reliability of the data analysis, an external usability expert recoded a subset of the data independently, an approach suggested by Gray and Salzman (1998). According to intercoder-reliability guidelines, 10% of the comments were recoded 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 1: Number of Comments Elicited by Questionnaires and Interviews, Preceded by Thinking-Aloud or Not

	Comment Type	*Thinking-Aloud First*	*No Thinking-Aloud First*
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)
	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	Positive	5.38 (2.39)	7.50 (2.14)
	Negative	5.62 (3.02)	3.62 (2.77)
	Neutral	3.75 (1.98)	2.50 (1.51)
	Specific usability issues	12.00 (3.78)	11.38 (3.54)
	Appreciation of personalization	1.38 (.74)	1.62 (.74)
	Perceived relevance of search results	1.38 (1.06)	.62 (.92)

Note. Standard deviations are in parentheses.

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 the questionnaires or 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 1. 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 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. This led to the following results:

- positive issues, $t(14) = .10$, $p = .93$, $d = .93$
- negative issues, $t(14) = -.31$, $p = .76$, $d = .77$
- neutral issues, $t(9.44) = -.56$, $p = .59$, $d = .64$
- usability issues, $t(14) = -.40$, $p = .70$, $d = .72$
- appreciation of personalization, $t(14) = .00$, $p = 1.00$, $d = .99$
- perceived relevance of search results, $t(14) = -.72$, $p = .49$, $d = .59$

The number of comments provided during a questionnaire preceded by thinking-aloud did not differ significantly from the number of comments elicited by questionnaires that were not preceded by thinking-aloud. All of the tests have reasonably large to very large effect sizes.

Next, we tested whether the number of comments of different types, elicited by interviews preceded by thinking-aloud or not, differed significantly:

- positive issues, $t(14) = -1.88$, $p = .08$, $d = .51$
- negative issues, $t(14) = 1.38$, $p = .19$, $d = .51$

- neutral issues, $t(14) = 1.42$, $p = .18$, $d = .52$
- usability issues, $t(14) = .34$, $p = .74$, $d = .75$
- appreciation of personalization, $t(14) = -.67$, $p = .51$, $d = .59$
- perceived relevance of search results, $t(14) = 1.51$, $p = .15$, $d = .52$

Again, we did not find a significant difference between any of the comment types. Effect sizes are medium to large. On the basis of these results, we conclude that thinking-aloud or not did not influence the number of comments given on the questionnaire or during the interview afterward.

Ultimately, 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.

5.3. 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.

Specific usability issues. If applicable, we determined the specific usability issue to which a comment could be attributed. Table 2 shows how many times a comment on each specific issue was made in each questionnaire, interview, or thinking-aloud session. The following is one example of a comment on the breadth of experience:

> 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:

> 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."

Analyses of variance (ANOVAs) were conducted to ascertain whether the number of comments yielded by each method differed. The ANOVAs uncovered significant differences for the number of comments on

- the usability issues of 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, $F(2, 45) = 6.80$, $p < .01$, $\omega = .52$.

Table 2: Number of Comments on Specific Issues for Personalization per Session, Elicited by Each Method

Method		*Predictability*	*Comprehensibility*	*Controllability*	*Unobtrusiveness*	*Privacy*	*Breadth of Experience*	*System Competence*
Questionnaire	*M*	1.00	2.19	0.75 (T)	1.06 (T)	1.50 (T)	1.44	1.00
	SD	0.37	0.75	0.45	0.68	0.63	0.89	0.63
Interview	*M*	1.19	2.19	1.13 (Q/T)	1.56 (T)	1.63 (T)	2.19 (T)	1.81 (T)
	SD	0.54	1.05	0.50	1.03	1.03	1.17	1.38
Thinking-aloud	*M*	0.69	1.50	0.06	0.06	0.06	1.00	0.89
	SD	0.87	1.32	0.25	0.25	0.25	0.97	0.50

Note. A letter behind a mean value means that the mean value for a method is significantly higher than the mean value of the method the letter corresponds with: Q = questionnaire; T = thinking-aloud.

No significant differences were found in the case of predictability, $F(2, 45) = 2.57$, $p > .05$, and comprehensibility, $F(2, 45) = 2.23$, $p > .05$. Table 2 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 that were displayed can be found in Table 2. 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, although thinking-aloud itself supplied only a marginal number of comments on the specific usability issues for personalization.

Appreciation of personalization and the perceived relevance of search results. Perceived relevance is the most important quality factor for personalized search engine output. Table 3 shows how many comments on the appreciation of personalization and perceived relevance of search results each method yielded. The following is one example of a comment on the appreciation of personalization:

> 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 3: Number of Comments on Appreciation of Personalization (Appr.) and Perceived Relevance of Search Results (Relevance) per Session, Elicited by Each Method

Method		*Appr.*	*Relevance*
Questionnaire	*M*	1.00 (T)	0.44
	SD	0.82	1.03
Interview	*M*	1.50 (T)	1.00
	SD	0.73	1.03
Thinking-aloud	*M*	0.13	6.13 (Q/I)
	SD	0.34	2.36

Note. A letter behind a mean value means that the mean value for a method is significantly higher than the mean value of the method the letter corresponds with: Q = questionnaire; I = interview; T = thinking-aloud.

ANOVAs 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 3. 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, questionnaires 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).

Positive, neutral or ambiguous, 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 4 shows the differences in the number of positive, neutral or ambiguous, or negative comments that were elicited. We performed ANOVAs to see whether the number of the differently valued comments each method yielded differed. There appeared to be significant differences in all cases, that is, 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 4. 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.

Problem severity. After analyzing the collected comments with a quantitative approach, we looked at their actual content. As the formative evaluation stage is focused on generating redesign input, we concentrate our efforts on negative comments only. By grouping and naming 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 displays each method's

Table 4: Number of Differently Valued Comments on Personalization per Session, Elicited by Each Method

Method		*Positive*	*Neutral*	*Negative*
Questionnaire	Mean	3.94 (T)	1.56	4.88
	SD	2.54	1.32	3.14
Interview	Mean	6.44 (Q/T)	3.13 (Q/T)	4.63
	SD	2.45	1.82	2.99
Thinking-aloud	Mean	0.81	1.00	8.31 (Q/I)
	SD	1.05	1.21	2.75

Note. A letter behind a mean value means that the mean value for a method is significantly higher than the mean value of the method the letter corresponds with: Q = questionnaire; I = interview; T = thinking-aloud.

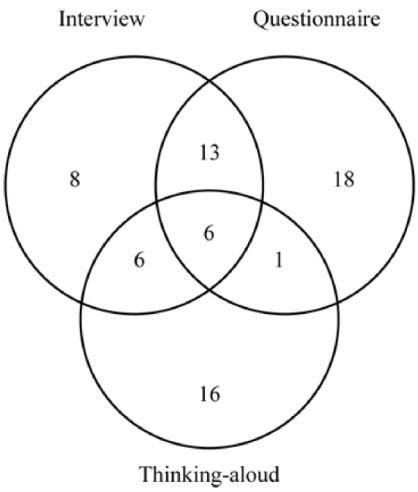

FIGURE 5 Problematic issues related to personalization uncovered by each method.

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.

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). In line with Høegh and Jensen (2008), Hornbæk and Frøkjær (2005), and Kjeldskov and Stage (2004), we classify problems as critical, serious, or minor. The following definitions are derived from Duh, Tan, and 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. The quality of coding was again assessed using an external usability expert who recoded the problems, according to the guidelines described before. A comparison of the original and recoded data set resulted in a Cohen's Kappa of .76, which stands for "good agreement."

We identified two critical, 13 serious, and 53 minor problems. 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 6 and 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. 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. Finally, each method produced a unique set of problems.

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.

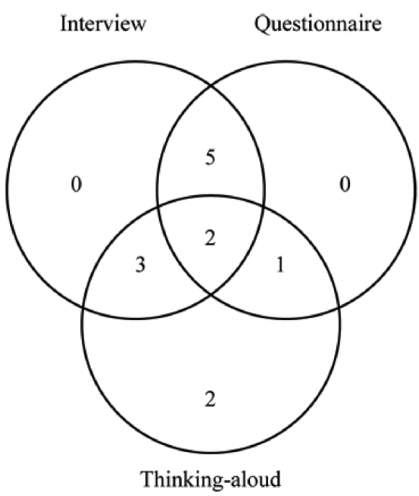

FIGURE 6 Serious problems relating to personalization uncovered by each method.

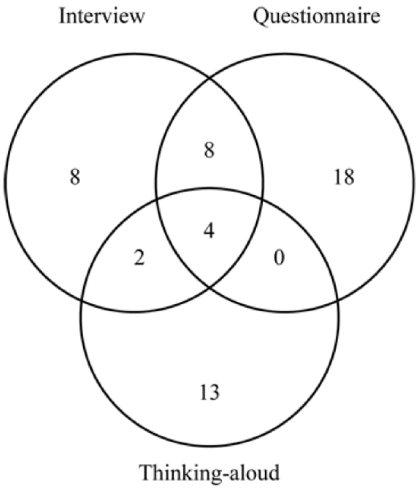

FIGURE 7 Minor problems relating to personalization uncovered by each method.

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. 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.

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 at .75. In total, there were 36 problems, of which 30 were minor, five were serious, and one was critical. The critical problem was brought forward by 12 participants

Table 5: 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

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 needs, characteristics, and context.

6. DISCUSSION AND CONCLUSIONS

6.1. Uncovering Specific Issues

Thinking-aloud has an important function in the iterative design process, 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, 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 upon), one cannot do without thinking-aloud, as it is the only method that unearths 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, and so on, are of a more general nature. Participants seem to be able to comment on them only when they are explicitly confronted with 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 usability issues for personalization than the questionnaire while both methods are, in this respect, superior to thinking-aloud. 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, comments elicited by the interview seem to be positively biased. It might be that participants want to save face (e.g., 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.

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 on 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.

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 personalization, but this does not affect the general trend over multiple methods. Examples of such similar forms of personalization include inserting or removing 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 layout of text. They also have a similar goal as they aim to guide the user to personally relevant information. Whether 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. 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.

6.4. The Integrated, Formative Evaluation of a Personalized System

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 or 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.

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).

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 or constructive interaction) 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.

6.5. Implications for Practitioners

This research has the following implications for practitioners who wish to evaluate a system with personalized presentation features in order to generate redesign input. First, one should use a combination of thinking-aloud and questionnaires. Thinking-aloud informs the evaluator about how the participants perceive the quality of personalization and their rationale behind this judgment. With this information, redesign can improve the personalization, thereby increasing system usability and usefulness. Furthermore, thinking-aloud will identify most of the critical and serious problems. Questionnaires were the best method for eliciting comments from the participant about specific usability issues for personalization. With this information, the user experience of the system can be improved upon. Second, "generic" usability issues also play a very important role in the redesign of a personalized system. Personalization can suffer from generic usability issues like unintuitive navigation. Therefore, eliciting comments on usability issues for personalization should be seen as a supplement to, and not as a substitute for, focusing on generic usability issues.

REFERENCES

Akoumianakis, D., Grammenos, D., & Stephanidis, C. (2001). User interface adaptation: evaluation perspectives. In C. Stephanidis (Ed.), *User interfaces for all* (pp. 339–352). Mahwah, NJ: Erlbaum.

Allwood, C. M., & Kalén, T. (1997). Evaluating and improving the usability of a user manual. *Behaviour & Information Technology*, *16*(1), 43–57.

Benbunan-Fich, R. (2001). Using protocol analysis to evaluate the usability of a commercial web site. *Information & Management*, *39*(1), 151–163.

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

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

Carroll, 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*, 123–135.

Carter, P. (2007). Liberating usability testing. *interactions*, *14*, 18–22.

Cordes, R. E. (2001). Task-selection bias: A case for user-defined tasks. *International Journal of Human–Computer Interaction*, *13*, 411–419.

De Jong, M., & Schellens, P. J. (2000). Toward a document evaluation methodology: What does research tell us about the validity and reliability of evaluation methods? *IEEE Transactions on Professional Communication*, *43*, 242–260.

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, UK: Springer-Verlag.

Doubleday, A., Ryan, M., Springett, M., & Sutcliffe, A. (1997). *A comparison of usability techniques for evaluating design*. Paper presented at the conference on Designing Interactive Systems: Processes, Practices, Methods, and Techniques, Amsterdam, the Netherlands, August 1997.

Duh, H. B., Tan, G. C. B., & Chen, V. H. (2006, September). *Usability evaluation for mobile device: A comparison of laboratory and field test*. Paper presented at MobileHCI, Helsinki, Finland.

Ebling, M. R., & John, B. E. (2000). *On the contributions of different empirical data in usability testing*. Paper presented at Designing Interactive Systems: Processes, Practices, Methods, and Techniques, New York, NY, August 2000.

Fossey, E., Harvey, C., McDermott, F., & Davidson, L. (2002). Understanding and evaluating qualitative research. *Australian & New Zealand Journal of Psychiatry*, *36*, 717–733.

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, Germany: Springer-Verlag.

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

Hartson, H. R., Andre, T. S., & Williges, R. C. (2001). Criteria for evaluating usability evaluation methods. *International Journal of Human–Computer Interaction*, *13*, 373–410.

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*, 2030–2044.

Hertzum, M., Hansen, K. D., & Andersen, H. H. K. (2009). Scrutinising usability evaluation: does thinking aloud affect behaviour and mental workload? *Behaviour & Information Technology*, *28*, 165–181.

Høegh, R. T., & Jensen, J. J. (2008). A case study of three software projects: can software developers anticipate the usability problems in their software? *Behaviour & Information Technology*, *27*, 307–312.

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 Intelligent User Interfaces, Orlando, FL, January 1997.

Hoppmann, T. K. (2009). Examining the "point of frustration." The think-aloud method applied to online search tasks. *Quality and Quantity*, *43*, 211–224.

Hornbæk, K. (2010). Dogmas in the assessment of usability evaluation methods. *Behaviour & Information Technology*, *29*(1), 97–111.

Hornbæk, K., & Frøkjær, E. (2005, April). *Comparing usability problems and redesign proposals as input to practical systems development*. Paper presented at the Computer Human Interaction conference, Portland, OR.

Jameson, A. (2007). Adaptive interfaces and agents. In J. A. Jacko & A. Sears (Eds.), *Human–computer interaction handbook* (2nd ed., pp. 433–458). Mahwah, NJ: Erlbaum.

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*, 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: Reaching Through Technology, New Orleans, LA.

Kaufman, J. (2006). Practical usability testing [Electronic version]. *Digital web magazine*. Retrieved 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.), *ITS 2000* (LNCS 1839 ed., pp. 19–30). Berlin, Germany: Springer-Verlag.

Kjeldskov, J., & Stage, J. (2004). New techniques for usability evaluation of mobile systems. *International Journal of Human–Computer Studies*, *60*, 599–620.

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.

Kushniruk, A. W., & Patel, V. L. (2004). Cognitive and usability engineering methods for the evaluation of clinical information systems. *Journal of Biomedical Informatics*, *37*(1), 56–76.

Labaw, P. J. (1981). *Advanced questionnaire design*. Cambridge, UK: Abt books.

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*, 224–234.

Lindgaard, G. (1994). *Usability testing and system evaluation. A guide for designing useful computer systems*. London, UK: Chapman & Hall.

Miles, M. B., & Huberman, A. M. (1994). *Qualitative data analysis* (2nd ed.). Thousand Oaks, CA: Sage.

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.

Nielsen, J. (1998, October 4). Personalization is over-rated. *Alertbox*. Retrieved from http://www.useit.com/alertbox/981004.html

Nielsen, J., & Mack, R. L. (1994). *Usability inspection methods*. New York, NY: Wiley & Sons.

Patton, M. Q. (2002). *Qualitative research & evaluation methods* (3rd ed.). Thousand Oaks, CA: 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*, 482–493.

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

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, Hildesheim, Germany, October 2006.

Scott, D. B. (2008). Assessing text processing: A comparison of four methods. *Journal of Literacy Research*, *40*, 290–316.

Spool, J. M., Scanlon, T., Schroeder, W., Snyder, C., & DeAngelo, T. (1997). *Web site usability: A designer's guide*. North Andover, MA: User Interface Engineering.

Tsakonas, G., & Papatheodorou, C. (2006). Analysing and evaluating usefulness and usability in electronic information services. *Journal of Information Science*, *32*, 400–419.

Van der Geest, T. M. (2004). *Beyond accessibility: Comparing three web site usability test methods for people with impairments*. Paper presented at HCI2004: Design for Life, Leeds, England, September 2004.

Van Oostendorp, H., & De Mul, S. (1999). Learning by exploration: Thinking-aloud while exploring an information system. *Instructional Science*, *27*, 269–284.

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*, 261–281.

Weibelzahl, S. (2005). Problems and pitfalls in evaluating adaptive systems. In S. Chen & G. Magoulas (Eds.), *Adaptable and adaptive hypermedia systems* (pp. 285–299). Hershey, PA: IRM Press.

Zabed Ahmed, S. M. (2008). A comparison of usability techniques for evaluating information retrieval system interfaces. *Performance Measurement and Metrics*, *9*(1), 48–58.

APPENDIX

Questionnaire and Interview Items (Translated from Dutch)

- 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?