From: AAAI Technical Report SS-98-02. Compilation copyright © 1998, AAAI (www.aaai.org). All rights reserved.

Don't Make That Face:
a report on anthropomorphizing an interface

Alan Wexelblat

MIT Media Lab
20 Ames St
Cambridge, Massachusetts 02139
wex@media.mit.edu

Abstract

A study was performed, comparing two identical systems, which were equipped with different interfaces. One was a conventional data interaction, written in "standard" terse neutral English. The other was an anthropomorphized, conversational style interface. Users performed a set of tasks and gave feedback on the effect and affect of the systems. Our hypotheses were (1) that the affective measures would be improved in the conversational situation but (2) the standard interface would score higher on effective measures. The first hypothesis was weakly supported by the results; however, there was no support for the second hypothesis.

Introduction[1]

Don't anthropomorphize computers; they hate that!

We describe an experiment constructed to test what effect anthropomorphism in an interface might have. In particular, the goal was to build systems which were functionally identical but which had different interfaces so that they could be compared head-to-head.

It has long been a tenet of user interface design that anthropomorphism is unfailingly a bad idea. Proponents of direct manipulation design, such as Shneiderman, argue that interfaces should be solely designed to reflect the commands available and the objects that the user can change. Anthropomorphism in the interface is derided as inefficient and confusing (Lanier 1995, Shneiderman 1995).

Anthropomorphism here refers to the assumption by the interface or system of intentional language (Dennet 1989) or human-like characteristics such as desires. For example, a dialog message that referred to the computer (or program) in the first person -- "I will save your file" or "I need the next disk" -- would be anthropomorphic in our sense. We differentiate anthropomorphism from personalization; the latter is the natural human tendency to use intentional shortcuts when talking about inanimate objects. For example, we might say "my car knows how to drive itself home" or "this computer hates me." In these cases, we do not really believe that the car has knowledge or that the computer has the emotion of hatred; rather, we use these phrases as conversational shortcuts and we use them even while operating conventional direct-manipulation interfaces. Anthropomorphism involves the computer using human aspects such as intentions or appearances to facilitate a conversational or delegative style of interface.

Additionally, the use of human-seeming interface representations -- a visual embodiment of an interface agent -- is anthropomorphic, since the representation takes on human qualities. In the experiment reported in this paper, both techniques were employed in one of the systems -- called the anthro interface – and excluded in the other, called the minimal interface.

The goal of the experiment was to test two related hypotheses. The first was that the anthro system would score higher on measure of user affect, such as likability and friendliness. However, because the anthro system involved reading full intentional sentences instead of simple direct commands, it was also hypothesized that the minimal system would score higher on effectiveness measures such as ease of use and quality of results.

As will be explained in the body of the paper, the experiment produced some evidence for the first hypothesis but, surprisingly, no evidence for the converse hypothesis was found.

Related Work

Although there have been numerous studies of many varieties of direct manipulation interfaces (see Shneiderman (Shneiderman 1997) for a discussion of direct-manipulation interface design) there are no specific comparisons that we could find in the literature in which a direct manipulation system is compared one-on-one to an anthropomorphic system. The lack of directly comparable prior art is particularly true under the conditions used here, where the systems' functionality was held exactly constant.

[1]Copyright © 1998, American Association for Artificial Intelligence (www.aaai.org). All rights reserved.

Walker and her colleagues at DEC added an explicitly human-like face to an existing interface, and found that adding the human face did not improve the users' interactions with their system (Walker et al 1994). Their main finding was that the face was engaging but required more attention from the user, which may have been due to the fact that their face was animate and expressive. This kind of anthropomorphism likely required more effort to interpret than, for example, the static set of faces used by Kozierok in her calendar-scheduling agent, described in Maes' agents overview article (Maes 1994). Work by Takeuchi and his colleagues support the contention that realistic human facial animation in an interface is engaging but distracting from tasks (Takeuchi et al 1994).

Koda built a web-based system in which a user played draw poker against three computer agents represented by a variety of different faces ranging in complexity and expressiveness from smiley-face caricatures up through pictures of actual people (Koda 1996). Though the faces were not animated, they changed expression in response to events in the game, such as the agent players winning or losing a hand. Koda's experiments, like the one reported here, found no change in users' perception of the system's intelligence, as long as the representation was more complex than a simple line-drawing. King also found a similar effect, though his work dealt with comparisons of different facial representations outside an application context (King et al 1996).

Finally, Nass and Reeves at Stanford have been studying effects related to personification: they substitute computer systems into social psychological experiments and have repeatedly observed that subjects' reactions to these systems is significantly similar to their reaction to human partners in identical setups. For whatever reasons, people seem to treat and react to computer systems in much the same way they treat and react to humans (Nass and Reeves 1996).

The present experiment uses full cartoon representations of its agent instead of just faces, and the agents' expressions do not change, though different cartoons are used to represent different states of the system (e.g. searching for information versus presenting results). Thus it is most similar to Kozierok's work (Maes and Kozierok 1993), though she did not directly evaluate the effectiveness of her cartoon agent representations.

Experimental Setup

In order to eliminate confounding factors, the experiment used a setup where the interface was as divorced as possible from the functionality. The goal was to test only interface differences, while keeping all other aspects of the systems the same. The anthro and minimal interfaces were built as two series of web pages that gave users access to a back-end database. During the experiment, users entered data through the web pages and saw results as HTML files generated by Perl-language cgi-bin scripts. The Perl scripts and the database were identical for both conditions; only the HTML pages varied. We used the Netscape Navigator browser, version 3, running on an SGI workstation.

Figure 1 and Figure 2 show the main functional pages from each of the systems. The anthro system uses two characters, "Mr. Barnes" and "Fido." These characters are introduced to the user in the opening web pages of the system and at least one of them appears on every page shown or created for that system. They are the anthropomorphic representations of, and agents in, the system.

The text of the anthro system is written in a full-sentence, first-person style, as though "Mr. Barnes" was speaking to the user. For example, one of the options in the anthro system is labeled "I can show you the system-wide charts."

In the minimal system, no characters are used. The text makes explicit the fact that the user is interacting with a database and intentional language is avoided. For example, the same option as above appears in the minimal system as "System-Wide Charts." The minimal and anthro systems were identical in number of functions provided, in the groupings of functions, and in the ordering of function presentation in the interface.

The web pages served as interfaces to a database of information on musical artists. This database was originally constructed for the HOMR project by Metral (Metral 1995). The HOMR system used a collaborative filtering technique (Shardanand and Maes) to take in information -- in the form of numerical ratings -- and provide recommendations. The experimental systems we used accessed the same algorithms for both interfaces. Users completed tasks involving giving ratings and getting recommendations (the instructions given to subjects for completing the tasks are reproduced in Figure 3). They then filled out a questionnaire, which asked for their opinions on both the effect and affect of the systems.

Experiment

Twelve subjects participated in the experiment using one or the other of the systems. Subjects were recruited from the MIT campus by means of flyers, and included undergraduate and graduate students, staff members, and visitors. Subjects were paid US$5 for their participation, which lasted approximately 30 minutes. Six subjects were assigned to each condition, balanced so that equal numbers of male and female subjects used both systems. In addition, subjects were balanced by native language -- two-thirds of the users of both systems were native English speakers and one third had English as their second language. Subjects were all required to be proficient with the use of the Netscape Navigator web browser; they were also required to be unfamiliar with both the HOMR system and its commercial descendent, the Firefly system.

Subjects for both conditions were given the same set of directions, reproduced in Figure 3. They were told that the purpose of the experiment was to evaluate the quality of

the information (the recommendations) provided by the system. Subjects were not told until after they had filled out their post-test questionnaire that two alternative interfaces were being tested.

Subjects were repeatedly reassured during the experiment (both verbally by the experimenter and in the written instructions) that there were no right or wrong answers -- that their honest opinions were what mattered. The domain of music recommendation was felt to be appropriate because it is an area where people often have strong, easily expressed opinions. Since timing was not an issue, no practice sessions were done.

Subjects were given as much time as they wanted to complete the tasks. Subjects were not timed for the tasks, since the experiment was designed to involve reflection on the users' part and formulating opinions, which are not skill-based tasks. This is, however, typical of home-oriented entertainment applications that are expected to predominate in the next few years. The task itself involved finding the correct options within the system interfaces, selecting, them, and filling out HTML-based forms (for example, typing in a name and password, and selecting rating numbers from 1 to 7 on a popup widget).

As part of the post-task questionnaire, the subjects were asked to rate themselves in terms of two factors which it was felt might bias the comparison of systems: their musical knowledge and their expertise in using web-based systems. Our first fear was that users who were less knowledgeable about music would find the results harder to interpret, biasing them toward the anthro system. Our second fear was that users who were more expert at using web systems would be bored by the anthro presentation and would thus be biased toward the minimal system.

In our test population, users' reported levels of musical knowledge were not significantly different between the two groups (mean of 3.83 for the anthro group, 3.67 for the minimal group, $p > .10$). Users' mean level of expertise was slightly significantly different (means of 3.33 versus 2.83, $p < .10$); however, since this difference meant that more expert users ended up in the anthro category, there was no evidence to suggest that either of our initial fears came true.

Results

As shown in Figure 4, users rated the systems on a scale of 1 to 5 for the first part of the questionnaire, with 1 being the lowest value and 5 being the highest. These ratings compared the system used by the subject (anthro or minimal) against other Web-based systems they had used in the past. The first four questions resulted in no significant difference in their means ($p > .10$ on a standard t test in all cases). For the question "More useful," the mean for the anthro group was 3.67, and 3.5 for the minimal. For the question "Easier to use" the anthro mean was 3.67, the minimal was 3.83. For the question "Friendlier" the anthro mean was 3.83, and the minimal mean was 3.67. For the question "Smarter" the anthro and minimal means were both 3.5.

The question "More enjoyable" did produce a significant difference ($p < .05$) with the anthro mean being 4.17 and the minimal mean being 3.5.

The second set of questions was phrased as assertions about the system used by the subject. A rating of 1 indicated strong disagreement with the assertion, and a rating of 5 indicated strong agreement. The first of these, "I got high-quality results," favored the minimal system (mean of 3.83, anthro mean of 3.5) but not at a statistically significant level.

The next question, "I liked using the system" produced a favorable difference for the anthro system (4.833 versus 3.67, $p < .05$). This was also the highest mean response to any question by any group.

The question "I felt I got what I expected" gave a slight advantage to the minimal system (mean of 4 versus 3.83) but this was not statistically significant. Similarly, on the "I understood what the system would do" question, the minimal system received a mean rating of 3.83, which was better than the anthro system's 3.33, but not significantly so.

Discussion

As noted earlier, we were seeking evidence related to two hypotheses concerning the effect -- or quality -- of the two systems and their affect -- or user experience. In the real world, of course, these two dimensions are not completely separable. The quality of the experience we have inevitably affects our perception of the quality of the results (Picard 1997). However, this argument does not necessarily favor one system over another in our situation. If people get what they expect and find direct manipulation-style interfaces easier to use, then the minimal interface should benefit, just as the anthro system should benefit if users found it to be more enjoyable. It seems reasonable to hypothesize that the minimal system's advantage in the "expected" question is based on its similarity to common web-based systems that our subjects had seen before.

The questionnaire contained three queries related to the effectiveness of the system, three queries related to the affect, and three queries which attempted to see if the experience of using one system differed significantly from the experience of using the other. By our initial hypotheses, we expected the anthro system to score better on the affect questions (3, 4 and 5) and the minimal system to score better on the effect questions (1, 2, and 6). No pre-test hypotheses were formed about the quality questions (7, 8, and 9).

Unfortunately, the evidence is hardly unequivocal. The anthro system scored significantly better on only one measure, the enjoyability. The minimal system did not score significantly better on any of the effectiveness measures. In considering the impact of this failure, it should be kept in mind that these are not direct measures

of the effectiveness of the system, but rather are measures of the users' different perception of the effectiveness of direct manipulation versus anthropomorphized interfaces. However, if the slightly more experienced users in this study did not find the anthro system less effective, that can be seen as an indication that anthropomorphism of the sort studied here would not be likely to negatively affect quantitative measures of effectiveness.

Of the three quality measures, only the first, enjoyment, showed any significant difference, with the anthro system being rated more enjoyable. This is in line with the other studies in this area and is not particularly surprising, given that the anthropomorphized system used cartoon characters designed to be engaging.

This experiment and Koda's suggest that, even when dealing with an extremely realistic human interface representation, people are not fooled into thinking the system is more intelligent or more capable. This has been one of the most often-repeated concerns about anthropomorphized interfaces, but up to now there has not been a direct comparative test.

It is hard to count a lack of evidence as evidence against a hypothesis, so while we can say that the second major hypothesis of this experiment -- that the minimal system would be more effective -- was not supported, we cannot say that it was disproven in the general case.

There are two alternative explanations for the results related here. One is that the experiment is itself somehow flawed. A different experiment could be designed, which would measure other factors, such as possible impacts of anthropomorphized interfaces on users' performance of skilled tasks, on their acquisition or retention of learning, and so on. The explanation, then, would be that the experiment did not measure the important differences between the two styles. This sort of explanation is hard to discount without doing additional experiments.

However, the claims made against anthropomorphic interfaces have been so strident that it is surprising to see evidence that in fact an explicitly agented interface did not affect the users' ability to get quality information from the system. This leads to the other explanation, which is that the opponents of agent interfaces have been arguing with more passion than evidence and perhaps the burden should be on them to demonstrate conditions under which a significantly negative difference due to an anthropomorphic interface could be observed.

References

Dennett, Daniel. 1989. *The Intentional Stance.* Cambridge, Mass.: MIT Press.

King et al. 1996. The Representation of Agents: anthropomorphism, Agency, and Intelligence. In Proceedings of CHI'96. Reading, Mass.: Addison Wesley.

Koda, Tomoko. 1996. Agents with Faces: the Effects of Personification. In Proceedings of Human-Computer Interaction'96. London.

Lanier, Jaron. 1995. Agents of Alienation. *interactions* 2(3).

Maes, Pattie. 1994. Agents that Reduce Work and Information Overload. *Communications of the ACM* 37(7).

Maes, Pattie and Kozierok, Robin. 1993. Learning Interface Agents. In Proceedings of AAAI'93. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence, Inc.

Metral, Max. 1995. Motormouth: A Generic Engine for Large-Scale Real-Time Automated Collaborative Filtering. S.M. thesis., Program in Media Arts and Sciences, Massachusetts Institute of Technology.

Nass, Clifford and Reeves, Byron. 1996. *The Media Equation*, Cambridge, UK: Cambridge University Press.

Picard, Rosalind. 1997. *Affective Computing*, Cambridge, Mass.: MIT Press.

Shneiderman, Ben. 1997. *Designing the User Interface.* Reading, Mass.: Addison Wesley.

Shneiderman, Ben. 1995. Looking for the Bright Side of User Interface Agents. *interactions* 2(1).

Takeuchi et al. 1994. Situated Facial Displays Toward Social Interaction. In Proceedings of CHI'94. Reading, Mass.: Addison Wesley.

Walker et al. 1994. Using a Human Face in an Interface. In Proceedings of CHI'94. Reading, Mass.: Addison Wesley.

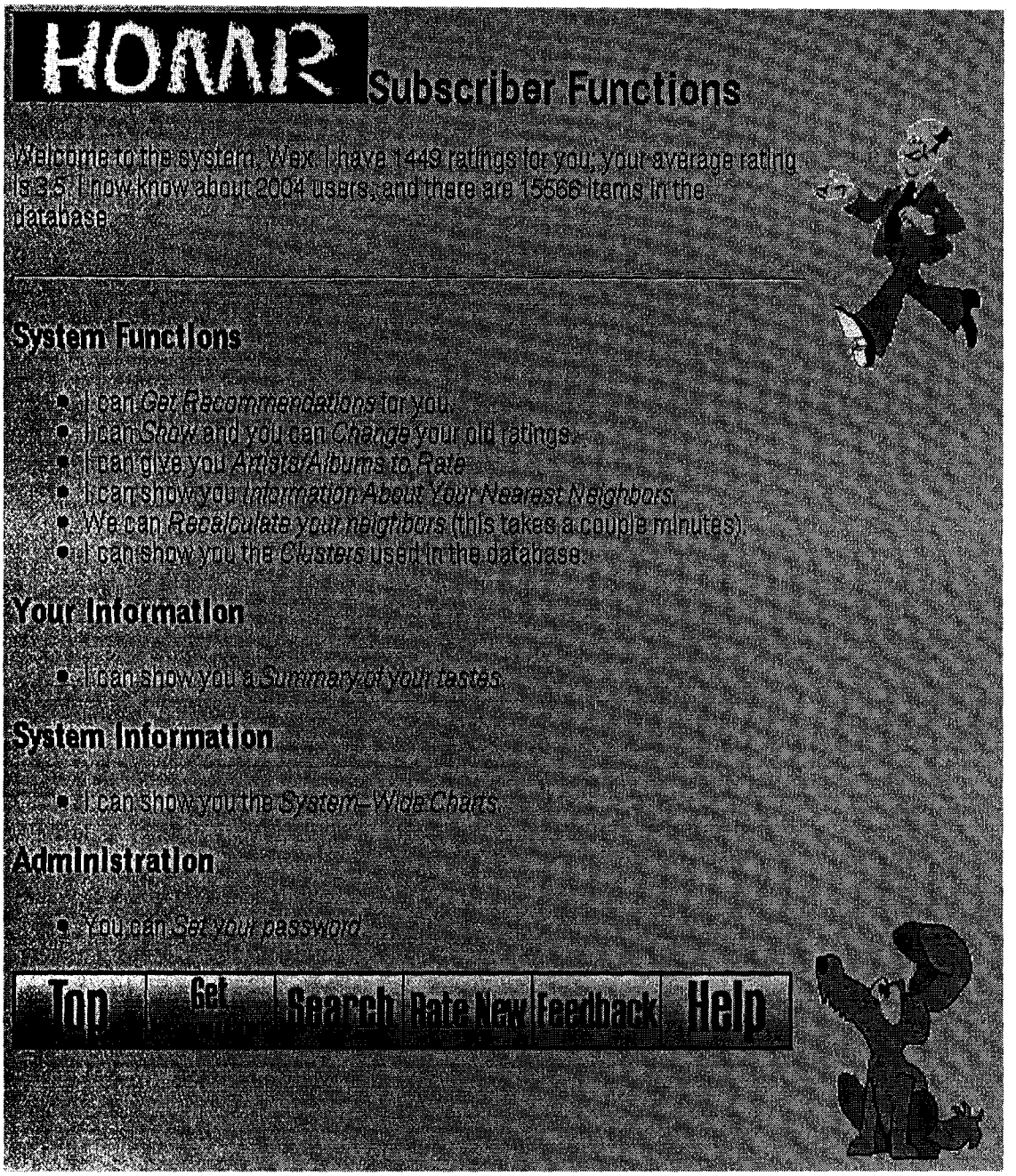

Figure 1: Anthro system subscriber functions page

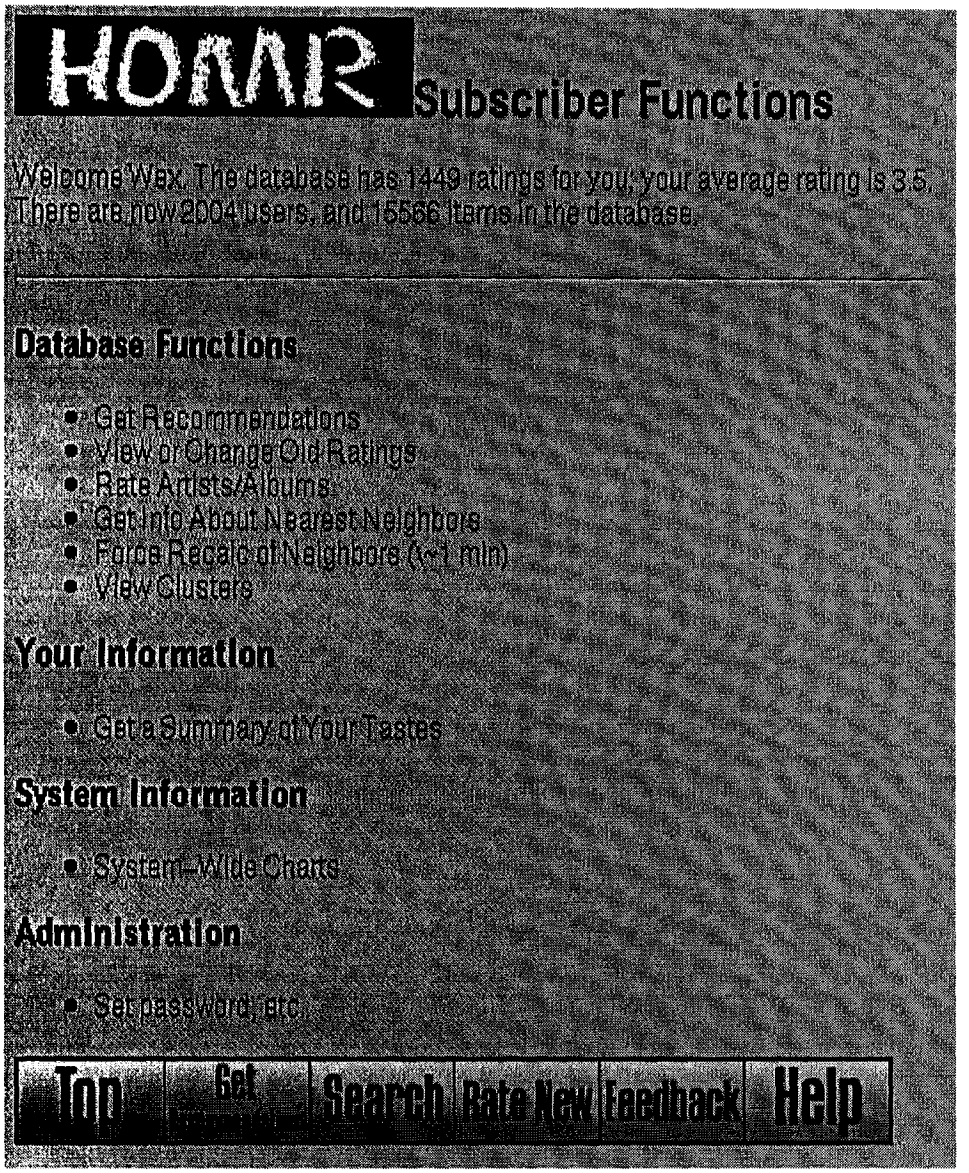

Figure 2: Minimal system subscriber functions page

Instructions Given to Subjects

Here are the steps you should follow when interacting with the system. In some cases the instructions may only make sense when you are looking at the particular web page. Remember that the system's "Help" functions have been turned off, so if you have any questions, please ask the experimenter.

1. Subscribe to the database.
2. Go to the Subscriber Functions.
3. Get some Artists/Albums to rate.
   Please use the initial survey method.
4. Rate as many of these things as you would like, but at least ten, please.
5. Submit your ratings.
6. Go to the Top level
7. View the System Chart for Highest/Lowest rated artists.
8. Go to the Top level.
9. Look at your old ratings.
10. Make some new ratings.
    Please rate some of the Most-rated artists.
11. Submit your ratings.

Figure 3: Subject Instructions

Questionnaire Filled out by Subjects

On a scale of 1 to 5 where 1 is the lowest or least and 5 is the highest or most, please rate yourself and then give us your opinion about the system:

	Lowest				Highest
My level of music knowledge	1	2	3	4	5
My level of Web expertise	1	2	3	4	5

Compared to other web-based systems I have used, this one is:

	Lowest				Highest
More useful	1	2	3	4	5
Easier to use	1	2	3	4	5
Friendlier	1	2	3	4	5
Smarter	1	2	3	4	5
More enjoyable	1	2	3	4	5

For these questions, please answer on a scale of 1 to 5 where 1 indicates that you disagree strongly with the statement and 5 indicates that you agree strongly.

	Disagree				Agree
I got high-quality results	1	2	3	4	5
I liked using the system	1	2	3	4	5
I felt I got what I expected	1	2	3	4	5
I understood what the system would do	1	2	3	4	5

Figure 4: Post-Test Questionnaire