How the Interface Design Influences Users' Spontaneous Trustworthiness Evaluations of Web Search Results: Comparing a List and a Grid Interface

Yvonne Kammerer*
Knowledge Media Research Center, Tuebingen, Germany

Peter Gerjets†
Knowledge Media Research Center, Tuebingen, Germany

Abstract

This study examined to what extent users spontaneously evaluate the trustworthiness of Web search results presented by a search engine. For this purpose, a methodological paradigm was used in which the trustworthiness order of search results was experimentally manipulated by presenting search results on a search engine results page (SERP) either in a descending or ascending trustworthiness order. Moreover, a standard list format was compared to a grid format in order to examine the impact of the search results interface on Web users' evaluation processes. In an experiment addressing a controversial medical topic, 80 participants were assigned to one of four conditions with trustworthiness order (descending vs. ascending) and search results interface (list vs. grid) varied as between-subjects factors. In order to investigate participants' evaluation processes their eye movements and mouse clicks were captured during Web search. Results revealed that a list interface caused more homogenous and more linear viewing sequences on SERPs than a grid interface. Furthermore, when using a list interface most attention was given to the search results on top of the list. In contrast, with a grid interface nearly all search results on a SERP were attended to equivalently long. Consequently, in the ascending trustworthiness order participants using a list interface attended significantly longer to the least trustworthy search results and selected the most trustworthy search results significantly less often than participants using a grid interface. Thus, the presentation of Web search results by means of a grid interface seems to support users in their selection of trustworthy information sources.

CR Categories: H.3.3 [Information Search and Retrieval]: Search process; Selection process H.5.2 [Information Interfaces and Presentation (e.g., HCI)]: User Interfaces - Screen design; Evaluation/methodology

Keywords: WWW search, search engines, eye tracking methodology, evaluation processes, trustworthiness, viewing sequences, search results interface

1 Introduction

In recent years, the World Wide Web (WWW) has evolved into one of the most important information sources, offering easy access to vast amounts of information. Particularly for domains of personal concern such as medicine and healthcare using the Web has achieved great popularity [Eysenbach et al. 2002; Fox 2006; Morahan-Martin 2004]. When looking for online medical and health information, most Web users start by using a general search engine such as Google, Yahoo!, or MSN [Fox 2006; Moharan-Martin 2004]. As a response to the search terms entered by the user, these search engines usually return a rank-ordered list of search results, with the (hypothetically) most relevant and most popular Web pages being the highest-ranked ones [cf. Cho and Roy 2004]. For each search result on the search engine results page (SERP) a title, an excerpt of the content of the Web page, and its URL are displayed.

Many people who use the Web to retrieve medical and health information are not looking for a specific fact but for complex and comprehensive information, for instance, about the risks and benefits of specific medical treatments [cf. Moharan-Martin 2004]. In other words, medical Web search often goes far beyond simple fact-finding and involves rather complex information gathering. Moreover, it must be taken into account that - for instance in the case of looking for medical treatments - the information retrieved from the Web might strongly influence patients' decisions on which medical treatment to choose or to refuse [Fox 2006]. Considering the potential influence of Web information on important personal decisions, the trustworthiness of information sources becomes a pivotal issue.

However, as anyone can publish virtually any information on the Web, the WWW is characterized by a high heterogeneity of information sources differing, for instance, with regard to Web authors' expertise and motives. As a result, the trustworthiness of medical and health information available online varies considerably, with many Web sites containing misleading or even wrong information [Eysenbach et al. 2002]. Despite this fact, Web pages from different types of Web authors (e.g., scientific and other institutions, journalists, lay people, and companies) are usually interspersed in the results lists returned by search engines. Moreover, it is often the case that popular commercial Websites fit exactly to the search terms entered by the user, and thus are listed among the highest-ranked search results. However, due to the commercial interests of these information providers, the information on commercial Websites might easily turn out to be biased and onesided. As a consequence, premature or even wrong decisions may result.

Thus, in order to avoid the selection and use of incomplete, biased, or even false information, Web users are required to critically evaluate the search results by themselves in terms of trustworthiness [cf. Taraborelli 2008] - especially when dealing with controversial issues such as the effectiveness of specific medical treatments.

The focus of the present study was to investigate to what extent users spontaneously (i.e., without receiving any prompts) evaluate the trustworthiness of Web search results presented to them by a search engine when searching information about complex medical problems. Moreover, we examined the impact of the search results interface on Web users' spontaneous evaluation processes by comparing a list interface to a grid interface for presenting search results. Whereas the standard search results interface used by popular search engines such as Google, Yahoo!, and MSN/Bing

*e-mail: y.kammerer@iwm-kmrc.de
†e-mail: p.gerjets@iwm-kmrc.de

ETRA 2010, Austin, TX, March 22 – 24, 2010.
© 2010 ACM 978-1-60558-994-7/10/0003 $10.00

presents search results in a one-dimensional list format, recently some search engines have emerged which use a grid interface displaying search results in a two-dimensional layout [e.g., www.viewzi.com, www.gceel.com].

In the current study, we used eye tracking methodology to investigate Web users' evaluation processes on SERPs in addition to logging overt interactions with the search interface (e.g., the selection of search results). As eye movements reflect visual attention allocation [Duchowski 2007], eye tracking data allow reconstructing searchers' viewing patterns on a SERP, thereby revealing which search results they attended to, for how long, and in what order.

2 Related Work

2.1 Evaluating and Selecting Search Results

Several Web search studies have shown by means of eye tracking and/or log file analyses that users of popular search engines such as Google, Yahoo!, or MSN tend to read the search results from top to bottom [Cutrell and Guan 2007; Granka et al. 2004; Joachims et al. 2005; Salmerón et al. 2009]. Moreover, searchers spend most attention to the search results on top of a SERP and also predominantly select these first few links [Granka et al. 2004; Cutrell and Guan 2007; Guan and Cutrell 2007; Eysenbach and Köhler 2002; Joachims et al. 2005; Pan et al. 2007].

In order to investigate whether users' focus on the top search results was based on the ranking position of the search results, on users' relevance evaluations of the search results, or on a combination of both, Keane et al. [2008] and Pan et al. [2007] used a methodological paradigm in which the relevance order of search results on a Google SERP was experimentally manipulated. Search results were either presented in a regular order (descending relevance order) or in a systematically reversed order (ascending relevance order) with the first search result being the (hypothetically) least relevant one on the SERP. Results of these studies demonstrated that when the top search results on a SERP were the least relevant ones, participants viewed more search results than when the top search results were the most relevant ones. Furthermore, in this case, they sometimes selected a highly relevant search result placed further down the list. However, participants generally still paid most attention to the search results on top of the SERP and selected these results most often. Similarly, a study by Guan and Cutrell [2007] that varied the position of the target search result (i.e., the most relevant result) in a MSN SERP, showed that when the target result was placed lower on the SERP, participants still tended to click on one of the top search results, even though they viewed more search results than when the target search result was among the top positions on the SERP. These empirical findings indicate that searchers, although they seem to employ some degree of evaluations regarding the topical relevance of the search results presented on a SERP, often tend to degrade their own evaluations and rather obey the ranking determined by the search engine.

With regard to users' spontaneous trustworthiness evaluations of Web search results, which are in the focus of the current study, previous empirical findings, however, show an even less optimistic picture: Analyses of verbal protocols from participants engaged in complex search tasks revealed that the majority of participants' utterances referred to the topical relevance of search results. In contrast, evaluation criteria with regard to the expected trustworthiness of information sources (e.g., Website reliability, potential bias of information, authors' expertise and motives) were uttered only rarely [Brand-Gruwel et al. 2008; Gerjets et al. in press; Walraven et al. 2009]. These findings are in line with results from other Web search studies which found that lay people often do not spontaneously verify the trustworthiness of information obtained via the WWW, for instance, when searching for medical information [e.g., Eysenbach and Köhler 2002; Metzger et al. 2003; Fox 2006].

In summary, previous empirical findings indicate that Web users heavily rely on the ranking offered by search engines. They predominantly select highest-ranked search results and only rarely evaluate the trustworthiness of sources spontaneously. When searching for simple and uncontroversial facts this strategy of quickly selecting one of the top search results on a SERP can be considered highly efficient, as search engines' ranking algorithms usually work well for this type of task. However, when it comes to more complex information search tasks, such as making informed decisions about a specific medical treatment, the focus on selecting the highest-ranked search results might lead to a biased and incomplete view of the facts due to the heterogeneity of information sources on the Web. Therefore, especially when searching for complex and controversial contents on the Web, a critical evaluation of the trustworthiness of search results is crucial for making well-informed decisions. In the current study we investigated how changes in the design of the search results interface might influence whether users spontaneously engage in trustworthiness evaluations.

2.2 Alternative Search Results Interfaces

The studies reported in the last section addressed Web users' evaluation and selection behavior for conventional search results interfaces, presenting search results as a one-dimensional rank-ordered list. The studies yielded that such a list format sets a strong focus on the top search results and imposes a strict and non-ambiguous order in which to read and select search results. Searchers might thereby be inhibited to trust their own search results evaluations, or to engage in evaluations at all. As a consequence, they seem to be discouraged to select potentially more suitable search results displayed further down the list.

Thus, the question arises, whether changes in the format of the search results interface may reduce the impact of the ranking order, increase the awareness of the selection process, and thereby stimulate users to evaluate search results, for instance in terms of trustworthiness. A recent study by Salmerón et al. [2009] showed that the graphical-overview search results interface Kartoo [www.kartoo.com] supported a more free exploration and selection of search results across the SERP compared to a regular Google-like list interface. Resnick et al. [2001] tested a conventional list search results interface against a tabular interface in which the columns of a table corresponded to the different elements of the search results (title, excerpt, URL, and metadata). Results showed that the tabular interface supported a wider variety of search strategies than the conventional list interface.

Moreover, as mentioned earlier, some recent search engines use a grid format displaying search results in a two-dimensional layout. In this grid format search results are presented in multiple rows and columns. This implies that there is no strict and non-ambiguous order in which to read and select search results. It is unclear whether the ranking within a SERP is aligned horizontally (i.e., line-by-line, according to the regular western reading direction) or vertically (i.e., column-by-column), or whether there exists a ranking at all. Therefore, in a grid interface the decision about the reading order is left open to the user and the ranking is less salient. We hypothesized that these features of a grid format might stimulate users to spontaneously evaluate search results, not

only in terms of relevance but also with regard to trustworthiness, and consequently to select the most suitable search results according to their own evaluations.

3 Aims and Hypotheses

The aim of the current study was to investigate the effects of a grid search results interface on Web users' evaluation and selection processes on SERPs containing search results of varying trustworthiness. We adapted the methodological paradigm used by Keane et al. [2008] and Pan et al. [2007] by experimentally manipulating the trustworthiness order of the search results on a SERP. Search results were presented either in a descending or ascending trustworthiness order (i.e., an order with the most trustworthy or the least trustworthy search result presented first, respectively). Furthermore, we compared two different search results interfaces: one interface presenting search results in a conventional list format, and an alternative interface presenting search results in a grid format.

Based on results from previous studies [e.g., Gerjets et al. in press] we expected that when using a conventional one-dimensional list interface individuals would hardly consider the trustworthiness of search results in their spontaneous evaluations. In contrast, we assumed that when using a two-dimensional grid interface searchers would be more conscious about their selection process because the order of the search results is less clear and the search results ranking is less salient. Consequently, they might spontaneously engage in trustworthiness evaluations when selecting search results. This effect was expected to become particularly evident for the ascending trustworthiness order, that is, when the top search results were of low trustworthiness.

Based on the assumption that the interface format of a SERP influences Web users' spontaneous evaluation processes, we derived the following specific hypotheses.

It was hypothesized that a list format provides a strong affordance for users to start reading at the top of the list. Therefore, we assumed that users of a list interface would homogeneously show rather linear viewing sequences from top to bottom when inspecting the search results available on SERPs [cf. Cutrell and Guan 2007; Granka et al. 2004; Joachims et al. 2005; Salmerón et al. 2009]. In contrast, we expected a grid interface to cause a more free exploration of the search results due to their less clear ordering on the SERP. This would result in less linear viewing patterns on SERPs, with higher variations across individuals as compared to a list interface.

Furthermore, we hypothesized, in line with previous findings [e.g., Pan et al. 2007], that in a list interface the majority of attention would be given to the top search results and that these search results would also be predominantly selected. As a consequence, in a list interface with ascending trustworthiness order the least trustworthy search results (on top of the list) would be predominantly attended to and selected, whereas in a list interface with descending trustworthiness order the most trustworthy search results would be predominantly attended to and selected. In contrast, for a grid interface, we expected that instead of a strong focus on only the first view search results, all search results on a SERP would be attended to a more equal degree. Furthermore, it was hypothesized that the most trustworthy search results would be predominantly selected not only in a descending trustworthiness order, but also in an ascending trustworthiness order, as the impact of the ranking on users' selection should be reduced with a grid interface. Thus, we hypothesized that when search results were presented in an ascending trustworthiness order, the most trustworthy search results would be selected more often with a grid interface than with a list interface.

4 Method

4.1 Participants

Eighty university students (17 male, 63 female; mean age 24.04 years) from different majors at a German university participated in this experiment for either course credit or payment. Pharmacy and medical students were excluded from the study. Participants had normal or corrected to normal vision. The majority of participants reported to have medium to high Web search experience and skills and no prior knowledge about the content domain of the study (Bechterew's disease). All participants reported to use Google as their primary search engine.

4.2 Apparatus

Eye movements were recorded during task processing by means of a Tobii 1750 remote eye tracking system (sampling rate: 50 Hz; gaze position accuracy: 0.5°) with IR-cameras built into a 17-inch monitor set to a resolution of 1280 x 1024 pixels. The Web stimulus recording mode of the ClearView 2.7.1 analysis software was used to capture not only the eye movements, but also task performance processes (including mouse operations). The minimum fixation duration was set to 80 milliseconds with a fixation radius of 30 pixels. The viewing distance between the participants and the screen was fixed to 65 cm, using a chinrest in order to prevent head movements. Before starting the eye movement recording, participants were calibrated on the eye tracking system using a nine-point calibration. Additionally, we applied a nine-point calibration validation to determine the tracking offset for each participant across the screen. Based on this data, the recorded gaze data were mathematically corrected posthoc for potential systematic offsets by means of an interpolation algorithm. The Web materials (see below) were displayed on a 17-inch computer screen and were presented with Microsoft Internet Explorer 7 using "larger" *Text Size View* to counterbalance the inaccuracy of the eye tracking system.

4.3 Task

The task used in the experiment was to seek information on the WWW about two competing therapies for Bechterew's disease ('Radon therapy' and 'Infliximab therapy'). The reason for choosing this medical topic was that it is complex and controversially discussed. Participants were confronted with a request from a fictitious friend, who was recently diagnosed with Bechterew's disease and therefore asked for advice about which of the two therapies to undergo. Participants were given eight minutes of time to conduct a Web research regarding the pros and cons of both therapies in order to make an informed decision and to give a recommendation to their friend.

4.4 Web Materials

For their Web research, participants were provided with two preselected SERPs, one for each therapy. Each of the two SERPs was accessible by means of a start Web page containing a brief description of the therapy and the task, as well as a hyperlink with the search terms used to generate the SERP. The search terms were the German words for "Bechterew's disease radon" and "Bechterew's disease infliximab". Each of the SERPs contained nine search results. For their Web research participants were given four minutes per SERP. Participants were not allowed to

generate new SERPs by changing the search terms. Participants could access all Web pages corresponding to the 18 search results presented. All Web pages were relevant to the search topic with regard to the content of the information provided. However, the collection of search results and Web pages for each of the two SERPs reflected the given heterogeneity of information sources on the Web, including Web pages provided by official institutions (e.g., department of health), industry and companies (e.g. health farms or pharma industry), and lay people (e.g. discussion pages). Furthermore, the Web pages contained partly conflicting information about pros and cons of the two therapies. To guarantee a standardized and controlled experimental setting, both the SERPs and the Web pages linked to them were put offline. All hyperlinks within the Web pages were disabled (except for the "back"-button of the browser to return to the SERP). Apart from the experimental manipulation of the interface (list versus grid, for details see below) the SERPs were displayed in Google style (cf. logo, font style and colors of search results, etc.) because of people's familiarity with this search engine. However, ads and the hyperlinks "in cache" and "similar pages" were not included on the SERPs.

4.5 Experimental Design

The experiment was a 2 (between-subjects) x 2 (between-subjects) x 9 (within-subjects) mixed-model factorial design. As a first factor the *trustworthiness order* of the search results was varied between subjects. Trustworthiness order was defined in a pilot-study where 24 participants were given two sets with nine search results each. Participants' task was to order the search results of a set according to the expected trustworthiness of the corresponding Web pages from 1 = most trustworthy to 9 = least trustworthy. Based on this data, two different trustworthiness orders were constructed for the SERPs: a *descending trustworthiness order*, with the most trustworthy search result presented first and the least trustworthy search result presented last, and a reversed *ascending trustworthiness order* with the least trustworthy search result presented first. For the grid interface (for details see below) search results were arranged line-by-line, that is, from left to right in each of the three rows, following the regular western reading direction. Participants were randomly assigned to one of the four conditions with 20 participants serving in each of the conditions.

As a second factor the *search results interface* was varied between subjects. The SERPs with nine search results were either presented as a standard *list interface* with search results listed from top to bottom or as a *grid interface* with search results arranged in three rows and three columns (see Figures 1 and 2).

Additionally, s*earch result trustworthiness* was considered as a third factor, with search results of *nine trustworthiness ranks* (from 1 = most trustworthy to 9 = least trustworthy), in order to investigate potential differences in users' evaluation and selection behavior with regard to search results of varying trustworthiness [cf. factor 'relevance rank' in Keane et al. 2008].

4.6 Procedure

Participants were tested in individual sessions of approximately one hour. Before starting with the experiment participants were asked to fill in a short computer-based questionnaire to provide some demographic data and personal data about their Web search experience as well as their familiarity with Bechterew's disease. Furthermore, they received some general instructions about the Web search experiment and were calibrated on the eye tracking system.

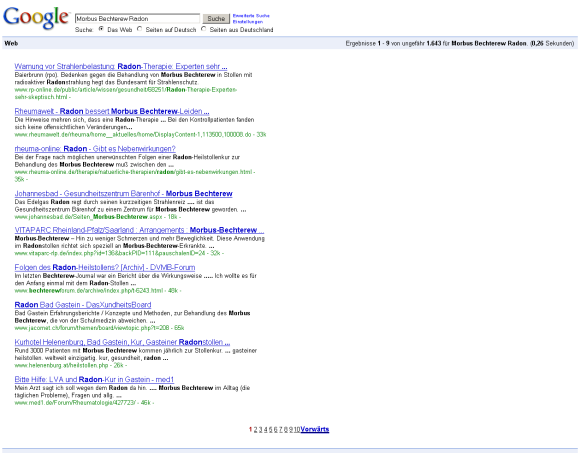

Figure 1. List interface.

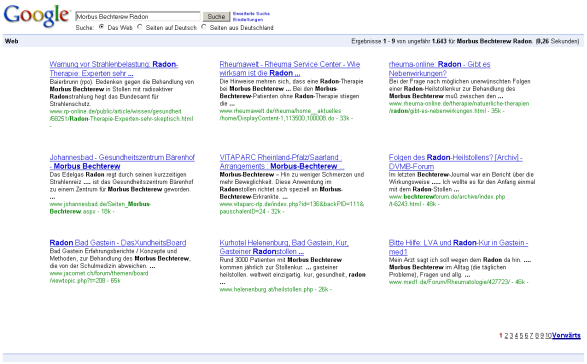

Figure 2. Grid interface.

Then, they underwent a training task for approximately two minutes to get acquainted with the experimental setup. This training task, which was about diet methods, was constructed equivalent to the subsequent main task. After the training task, participants received the instruction for the main task including the fictitious request of their friend. Subsequently, they were calibrated on the eye tracking system again and started their Web search regarding the first of the two therapies by clicking on the hyperlink on the start Web page leading to the first SERP. After four minutes, the information search regarding the first therapy was interrupted and a second start Web page was presented to the participants. By clicking on the hyperlink the second SERP appeared. To control for learning effects, the order of the two SERPs (related to the two therapies) was balanced among participants.

Eye movements and mouse clicks were captured during the entire eight minutes of task performance. Subsequent to the search task participants were asked to decide which of the two therapies they would recommend to their friend. However, to address our hypotheses only process measures during Web search were analyzed.

4.7 Dependent Measures

In order to investigate participants' evaluation processes on SERPs, we assessed their gaze behavior and their selection behavior with regard to the search results.

Gaze behavior. For the analysis of participants' gaze behavior so-called areas of interest (AOIs) were manually defined on the SERPs. AOIs are precisely specified areas on the screen for which eye tracking parameters (e.g., fixations) are aggregated. Each of the nine search results (including title, excerpt, and URL) was defined as a single polygonal AOI. Although the shape of the AOIs differed between the two search results interfaces (list vs. grid), their text content, font style, and font size were exactly the same for both interfaces (see Figures 1 and 2). As we did not expect any differences between the two SERPs (i.e., the 'Radon SERP' and the 'Infliximab SERP') gaze data of both SERPs were collapsed in the statistical analyses.

First, we analyzed participants' viewing sequences on a SERP, that is, the order in which the search results (i.e., the AOIs) were viewed (including returns to a SERP after having visited a Web page). More precisely, we measured the *homogeneity* and *linearity* of participants' viewing sequences in the four experimental conditions. This was done by means of the Levenshtein distance, a pairwise string-edit measure that calculates the edit distance between any two strings (e.g., viewing sequences). Edit distance is calculated as the minimum number of edit operations (insertions, deletions, or substitutions) needed to transfer one string into another [Josephon and Holmes 2002; Sankoff and Kruskal 1983]. Similar strings need fewer transformations and thus have smaller distances. Each participant provided one AOI string per SERP. For our analyses, an AOI string was derived from the first visits of the AOIs [cf. Joachims et al. 2005]. Thus, revisiting an already viewed AOI did not count as an element of the string. In order to assess the *homogeneity* of participants' viewing sequences within an experimental condition, Levenshtein distance of participants' viewing sequences was computed separately for each of the four conditions (by comparing each possible pair of strings within one condition). The edit distance calculated for a pair of strings was converted into a normalized similarity percentage (by dividing it by the length of the longer string and subtracting the result from 1). For each participant mean similarity percentages with the nineteen other participants were calculated. Finally, the mean similarity percentages in the different conditions were compared statistically.

Furthermore, in order to assess the *linearity* of participants' viewing sequences on a SERP within an experimental condition, we computed the similarity percentage between participants' string and a linear string (AOI 1, AOI 2, AOI 3, AOI 4, …, AOI 9). This linear string reflected a top-to-bottom sequence in a list interface and a line-by-line sequence in a grid interface, respectively. Participants' similarity percentages with the linear string in the different conditions were compared statistically. Additionally, as in a grid interface both a horizontal line-by-line sequence and a vertical column-by-column sequence are plausible, the similarity percentage between participants' string and a linear column-by-column string was computed (AOI 1, AOI 4, AOI 7, AOI 2, AOI 5, AOI 8, AOI 3, AOI 5, AOI 9). Finally, grid interface participants' similarity percentages with the column-by-column string as well as with the line-by-line string were compared statistically.

As a third dependent gaze measure the *total dwell time* (in milliseconds) on a search result of a specific trustworthiness rank (1-9) was measured, that is, the total time for which participants attended to a search result of a specific trustworthiness rank on a SERP.

Selection behavior. With regard to participants' selection behavior the search results participants selected in order to access a Web page were recorded. Search result selections for both SERPs (i.e., the 'Radon SERP' and the 'Infliximab SERP') were aggregated so that each search result of a specific trustworthiness rank could be selected zero to two times per person (i.e., re-openings of a Web page were not counted). As dependent measure, the *selection frequency* (i.e., 0-2 times) of a search result of a specific trustworthiness rank (1-9) was counted.

5 Results

Overall, participants spent 14.52% of their 8 minutes of Web search on SERPs on average. During this time they viewed nearly all (16.71) of the 18 search results and they selected 9.16 search results on average. ANOVAs (trustworthiness order x interface) showed no significant differences between the four experimental conditions with regard to the number of search results selected (trustworthiness order: $F < 1$; interface: $F < 1$; trustworthiness order x interface: $F(1, 76) = 1.17$, $p > .20$), the number of search results viewed (trustworthiness order: $F(1, 76) = 2.65$, $p > .10$; interface: $F(1, 76) = 1.60$, $p > .20$; trustworthiness order x interface: $F < 1$), or the proportion of time spent on SERPs and on Web pages (all Fs < 1).

5.1 Viewing Sequences

With regard to the *homogeneity* of participants' viewing sequences on a SERP, as measured by participants' mean similarity percentages provided by the Levenshtein algorithm, an ANOVA (trustworthiness order x interface) showed a significant main effect of interface ($F(1, 76) = 179.77$, $p < .01$). Viewing sequences of participants using a list interface were significantly more homogenous, with 61.21% similarity (descending: 60.20%; ascending: 62.23%), than were those of participants using a grid interface, with 38.26% similarity (descending: 39.05%; ascending: 37.46%). Besides this, there was neither a main effect of trustworthiness order ($F < 1$) nor a significant interaction between the two factors ($F(1, 76) = 1.13$, $p > .20$).

With regard to the *linearity* of participants' viewing sequences on a SERP, as measured by the similarity percentages of participants' string to a linear string (i.e., top-to bottom or line-by-line, respectively), the ANOVA also showed a significant main effect of interface ($F(1, 76) = 5.74$, $p < .05$). Viewing sequences of participants using a list interface had a similarity of 74.31% with the linear string (descending: 73.33%; ascending: 75.28%), whereas those of participants using the grid interface only had a similarity of 47.78% with the linear string (descending: 50.00%; ascending: 45.56%). Besides this, there was no main effect of trustworthiness order and no significant interaction between the two factors (both Fs < 1). Additionally, for the grid interface conditions the similarity percentage between participants' string and a linear column-by-column string was calculated. A repeated-measures ANOVA with the data of the 40 grid interface users showed that participants' viewing sequences had an even lower similarity of 34.44% with the column-by-column string (descending: 35.83%; ascending: 33.06%) than with the line-by-line string ($F(1, 38) = 10.32$, $p < .01$).

Furthermore, to analyze the heterogeneity of grid interface users' viewing sequences in greater detail, a hierarchical cluster analysis with the two variables 'line-by-line similarity percentage' and 'column-by-column similarity percentage' was conducted. The cluster analysis identified three different subgroups of grid interface users: The biggest group, comprising 28 participants (13 in the descending and 15 in ascending trustworthiness order), showed a rather low similarity percentage with both the line-by-line string ($M = 45.64\%$) and the column-by-column string

(M = 29.37%). The second group, comprising seven participants (descending: 4; ascending: 3), showed a moderate similarity with the column-by-column string (M = 59.52%). Finally, the third group, comprising five participants (descending: 3; ascending: 2) showed a high similarity percentage with the line-by-line string (M = 82.22%).

5.2 Total Dwell Time on a Search Result

A repeated-measures ANOVA (trustworthiness order x interface x search result trustworthiness) showed no significant main effects of trustworthiness order ($F(1, 76) = 1.12$, $p > .20$), interface ($F(1, 76) = 2.40$, $p > .10$), and search result trustworthiness ($F(4.84, 368.13) = 1.65$, $p > .10$) on the total dwell time on a search result. There were significant two-way interactions between trustworthiness order and interface ($F(1, 76) = 4.07$, $p = .05$) and between trustworthiness order and search result trustworthiness ($F(4.84, 368.13) = 18.50$, $p < .01$; Greenhouse-Geisser corrected). However, these interactions have to be interpreted in the light of a significant three-way interaction between all three factors ($F(4.84, 368.13) = 2.51$, $p < .05$; Greenhouse-Geisser corrected).

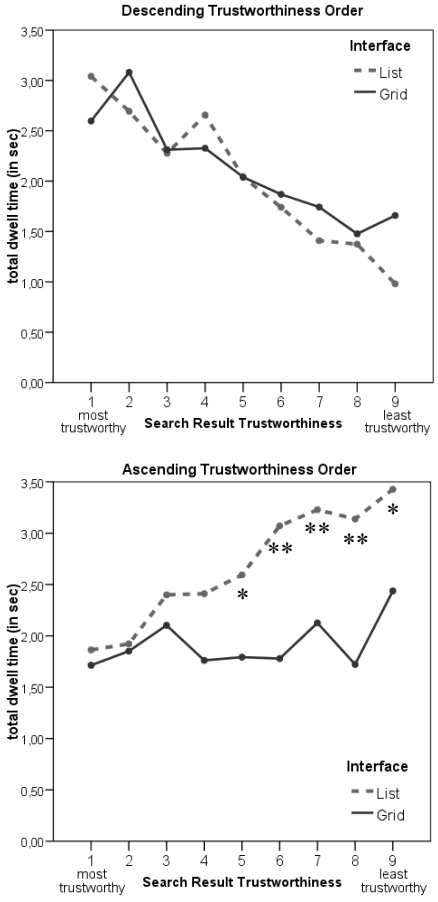

Figure 3. Total dwell time on a search result for descending (top) and ascending trustworthiness order (bottom).
(* denotes significant differences at 5%-level and ** at 1%-level)

Bonferroni-adjusted posthoc tests revealed that in the descending trustworthiness order there were no significant differences between the two interfaces with respect to participants' total dwell time on search results of any of the nine trustworthiness ranks (see Figure 3, top). In contrast, in the ascending trustworthiness order participants using the list interface attended significantly longer to the five least trustworthy search results than participants using the grid interface (see Figure 3, bottom). In other words, with a list interface the top search results (i.e., in the descending trustworthiness order the most trustworthy ones and in the ascending trustworthiness order the least trustworthy ones) received significantly more attention than the lower ones. In contrast, with a grid interface in both the descending and ascending trustworthiness order total dwell times on the search results of different trustworthiness ranks did not differ significantly (except for trustworthiness rank 2 which was attended to significantly longer than trustworthiness rank 6, 7, 8, and 9 in the descending trustworthiness order). That is, with a grid interface nearly all search results on a SERP were attended to equivalently long.

5.3 Frequency of Search Result Selection

A repeated-measures ANOVA (trustworthiness order x interface x search result trustworthiness) showed a significant main effect of search result trustworthiness ($F(8, 608) = 29.23$, $p < .01$) and a significant two-way interaction between trustworthiness order and search result trustworthiness ($F(8, 608) = 4.24$, $p < .01$) on the selection of a search result. However, these effects have again to be interpreted in the light of a significant three-way interaction between all three factors ($F(8, 608) = 2.24$, $p < .05$).

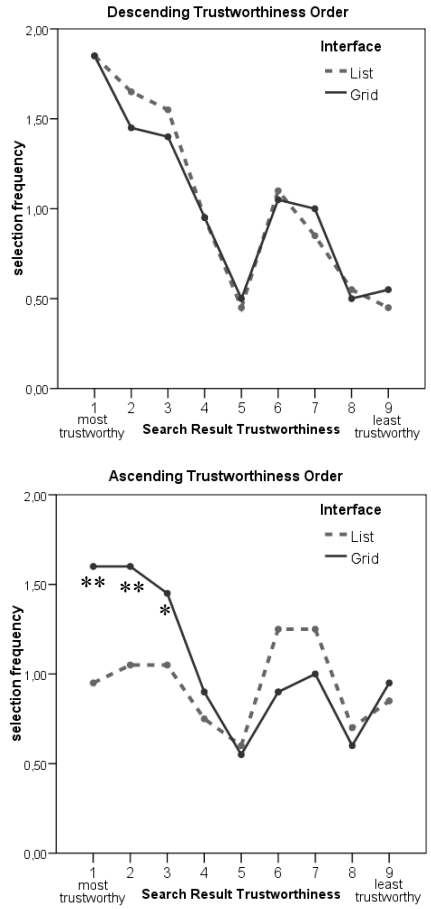

Figure 4. Frequency of search result selection for descending (top) and ascending trustworthiness order (bottom).
(* denotes significant differences at 5%-level and ** at 1%-level)

Bonferroni-adjusted posthoc tests revealed that - similar to the pattern for the total dwell time - in the descending trustworthiness order there were no significant differences between the two interfaces with respect to participants' selection frequency of search results of any of the nine trustworthiness ranks (see Figure 4, top). In contrast, in the ascending trustworthiness order participants using the grid interface selected the three most trustworthy search results significantly more often than participants using the list interface (see Figure 4, bottom). In other words, whereas with a list interface the three most trustworthy search results were selected significantly less often in the ascending trustworthiness order than in the descending trustworthiness order, with a grid interface there were no differences in search results selection with respect to trustworthiness order.

6 Discussion

The purpose of the current study was to investigate to what extent users spontaneously evaluate the trustworthiness of Web search results presented by a search engine when searching information about a complex problem, such as the effectiveness of specific medical treatments. Moreover, the study aimed at exploring the impact of the search results interface on users' evaluation processes. Therefore, we compared users' eye movements and mouse clicks with regard to search results of varying trustworthiness on a standard list interface and on a grid interface. First of all, our findings are consistent with results from previous Web search studies [cf. Cutrell and Guan 2007; Granka et al. 2004; Joachims et al. 2005; Salmerón et al. 2009] showing that users of a list interface quite homogeneously exhibited rather linear viewing sequences from top to bottom when inspecting the search results on a search engine results page (SERP).

In contrast, our analysis of eye tracking data showed that a grid interface caused more heterogeneous and less linear (i.e., line-by-line or column-by-column) viewing sequences on SERPs than a list interface. Similarly, a cluster analysis revealed that the majority of grid interface users showed neither linear line-by-line nor linear column-by-column viewing sequences on SERPs. This also explains the high heterogeneity of viewing sequences among grid interface users. To conclude, the study provides evidence that a grid search results interface supports a more free exploration of the search results than a list interface.

Second, also consistent with previous findings, the current study showed that when using a list interface the majority of attention was given to the top search results, whereas when using a grid interface, as expected, nearly all search results returned by the search engine were attended to equally long. Therefore, in line with our expectations, when the first search results were of low trustworthiness (ascending trustworthiness order), list interface users attended significantly longer to these less trustworthy search results than grid interface users. Moreover, in this case, list interface users selected the most trustworthy search results significantly less often than grid interface users. This suggests that the list format hinders Web users to deviate from the linear order and to select more trustworthy search results displayed at lower positions on the SERP. Furthermore, the findings seem to confirm our expectations that a grid interface stimulates Web users in complex search tasks to spontaneously evaluate the trustworthiness of search results and to select trustworthy information sources according to their own evaluations.

It has to be noted, however, that list interface users did not select the least trustworthy search results more often than grid interface users, when the first search results were of low trustworthiness. This might indicate that list interface users, contrary to our expectations, did employ at least some degree of trustworthiness evaluations. In order to find out whether in list interfaces the reduced selection of more trustworthy search results placed at the end of the list was due to participants' lack of trustworthiness evaluations, their confidence in the search engine's ranking, or a combination of both, in future research one might combine eye tracking with verbal reports [cf. Hansen 1991; Van Gog et al. 2005].

7 Conclusion and Future Work

In summary, the results of the current study suggest that redesigning the interface of search engines by displaying search results in a grid format instead of a list format seems to be a promising way to stimulate users to spontaneously evaluate the trustworthiness of search results and to support the selection and use of high-quality information.

We acknowledge, however, that our study has certain limitations. First of all, participants in the current study were all university students in their early to mid twenties, quite experienced in Web search activities. Moreover, they conducted their Web search in a lab setting on an artificially designed search task with a pre-defined search time of 8 minutes and a finite set of search results comprising only 18 search results. These study conditions limit the generalization of the results to a broader range of users and contexts. For instance, it is likely that due to the artificial settings participants in the current study viewed and selected more search results than they would naturally do. Furthermore, the heterogeneous viewing sequences of grid interface users might have at least partly resulted from an initial orienting response to the novel layout. Moreover, our study does not address individual differences between Web users, for instance, with regard to their personal search styles on SERPs [Aula et al. 2005] or their epistemological beliefs about the trustworthiness of information [e.g., Kammerer et al. 2009].

Second, even though the results of this study support the use of a grid search results interface for complex information search tasks, it is unclear, whether this would hold for simple fact finding tasks. One might assume that when searching for simple and uncontroversial facts, a grid format might hamper rather than improve the information search process because the quick selection of one of the highest-ranked search results on a SERP is usually highly efficient for this type of task. Furthermore, in the current study we did not assess any subjective measures such as users' satisfaction with the search results interfaces. Future studies should explore Web users' subjective satisfaction and their acceptance level with regard to grid interfaces as these measures will also play a critical role for the implementation of grid-like search results interfaces.

Finally, future studies may explore how the layout of the search results interface interacts with other features of the interface design aiming at improving searchers' trustworthiness evaluations on SERPs, such as the presentation of source categories or user ratings [Kammerer et al. 2009; Taraborelli 2008] or thumbnail images of Web pages [Woodruff et al. 2002].

In conclusion, this study provides interesting insights into how the interface design of a search engine influences users' spontaneous trustworthiness evaluations of Web search results. Nonetheless, future research is required to address the above mentioned limitations and to further expand our understanding of the cognitive processes involved in the evaluation of Web search results displayed in different interface designs.

References

AULA, A., MAJARANTA, P., and RÄIHÄ, K.-J. 2005. Eye-tracking reveals the personal styles for search result evaluation. In *Proceedings of Human-Computer Interaction.* INTERACT '05, 1058-1061.

BRAND-GRUWEL, S., VAN MEEUWEN, L., and VAN GOG, T. 2008. The use of evaluation criteria when searching the WWW: An eye-tracking study. In *Proceedings of the EARLI SIG Text and Graphics,* Tilburg, NL, 34-37.

CHO, J., and ROY, S. 2004. Impact of search engines on page popularity. In *Proceedings of the 13th international Conference on World Wide Web.* WWW '04. ACM, New York, NY, 20-29.

CUTRELL, E., and GUAN, Z. 2007. What are you looking for? An eye-tracking study of information usage in Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI '07. ACM, New York, NY, 407-416.

DUCHOWSKI, A.T. 2007. *Eye tracking methodology: Theory and practice*. Springer, London.

EYSENBACH, G., and KÖHLER, C. 2002. How do consumers search for and appraise health information on the World Wide Web? Qualitative studies using focus groups, usability tests, and in-depth interviews. *British Medical Journal, 324,* 573-577.

EYSENBACH, G. POWELL, J., KUSS, O., and SA, E.-R. 2002. Empirical studies assessing the quality of health information for consumers on the World Wide Web. A systematic review. *Journal of the American Medical Association, 287*, 2691-2700.

FOX, S. 2006. Online health search 2006. *Pew Internet & American Life Project*. Washington, DC. Available online at http://www.pewinternet.org/~/media//Files/Reports/2006/PIP_Online_Health_2006.pdf.pdf

GERJETS, P., KAMMERER, Y., and WERNER, B. in press. Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning & Instruction*.

GRANKA, L., JOACHIMS, T., and GAY, G. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th Annual ACM SIGIR Conference on Research and Development in Information Retrieval*. SIGIR '04. ACM, New York, NY, 478-479.

GUAN, Z., and CUTRELL, E. 2007. An eye tracking study of the effect of target rank on Web search. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems.* CHI '07. ACM, New York, NY, 417-420.

HANSEN, J.P. 1991. The use of eye mark recordings to support verbal retrospection in software testing. *Acta Psychologica, 76*, 31-49.

JOACHIMS, T., GRANKA, L., PAN, B., HEMBROOKE, H., and GAY, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in Information Retrieval.* SIGIR '05. ACM, New York, NY, 154-161.

JOSEPHON, S., and HOLMES, M.E. 2002. Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In *Proceedings of the 2002 symposium on Eye Tracking Research & Applications.* ETRA '02. ACM, New York, NY, 43-49.

KAMMERER, Y., WOLLNY, E., GERJETS, P., and SCHEITER, K. 2009. How authority-related epistemological beliefs and salience of source information influence the evaluation of web search results - An eye tracking study. In N. A. Taatgen, & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society*. Cognitive Science, Austin, TX, 2158-2163.

KEANE, M.T., O'BRIEN, M., and SMYTH, B. 2008. Are people biased in their use of search engines? In *Communications of the ACM*, *51,* 49-52.

METZGER, M.J., FLANAGIN, A.J., and ZWARUN, L. 2003. College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education*, *41*, 271-290.

MORAHAN-MARTIN, J.M. 2004. How internet users find, evaluate, and use online health information: A cross-cultural review. *CyberPsychology & Behavior, 7*, 497-510.

PAN, B., HEMBROOKE, H., JOACHIMS, T., LORIGO, L., GAY, G., and GRANKA, L. 2007. In Google we trust: users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*, article 3. Available online at http://jcmc.indiana.edu/vol12/issue3/pan.html.

RESNICK, M.L., MALDONADO, C.A., SANTOS, J.M., and LERGIER, R. 2001. Modeling on-line search behavior using alternative output structures. In *Proceedings of the Human Factors and Ergonomics Society 45th Annual Conference*, Minneapolis, MN, USA, 1166-1171.

SALMERÓN, L., GIL, L., BRÅTEN, I., and STRØMSØ, H.I. 2009. Comprehension effects of signalling relationships between documents in search engines. *Submitted for publication.*

SANKOFF, D., and KRUSKAL, J.B. 1983. *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison*. Addison-Wesley, Reading, MA.

TARABORELLI, D. 2008. How the Web is changing the way we trust. In A. Briggle, K. Waelbers, P.A.E. Brey (Eds.), *Current Issues in Computing and Philosophy*. IOS Press, Amsterdam.

VAN GOG, T., PAAS, F., VAN MERRIËNBOER, J.J.G., and WITTE, P. 2005. Uncovering the problem solving process: Cued retrospective reporting versus concurrent and retrospective reporting. *Journal of Experimental Psychology: Applied, 11*, 237-244.

WALRAVEN, A., BRAND-GRUWEL, S., and BOSHUIZEN, H.P.A. 2009. How students evaluate sources and information when searching the World Wide Web for information. *Computers & Education, 25*, 234-246.

WOODRUFF, A., ROSENHOLTZ, R., MORRISON, J., FAULRING, A., and PIROLLI., P. 2002. A comparison of the use of text summaries, plain thumbnails, and enhanced thumbnails for web search tasks. *Journal of the American Society for Information Science and Technology,* 53, 172-185.