Separating the Wheat from the Chaff:

The role of evaluation instructions, user characteristics, and the search interface in evaluating information quality during Web search on medical and health-related issues

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

Dipl.-Psych. Yvonne Kammerer

aus Esslingen am Neckar

Tübingen

2011

Tag der mündlichen Qualifikation: 18.05.2011

Dekan: Prof. Dr. Wolfgang Rosenstiehl

1. Berichterstatter: Prof. Dr. Peter Gerjets

2. Berichterstatter: Prof. Dr. Ivar Bråten

Acknowledgements

I would like to take this opportunity to thank everyone who has supported me on the way towards my PhD.

Specifically, I would like to thank

- ❖ my supervisor Peter Gerjets

- ❖ the members of my dissertation committee:
  Ivar Bråten, Ulrike Cress, Stephan Schwan, and Ulrich Trautwein

- ❖ the members of our research group "Hypermedia" and of the research group "Multimedia", as well as all my other current and former colleagues at KMRC and my research students

- ❖ my cooperation partner Lalo Salmerón

- ❖ my former flatmates Birgit and Halszka

- ❖ and last but not least Christian, my family, and my friends

Thank you!

Table of contents

List of figures

List of tables

1   Introduction and overview

Imagine you would like to inform yourself about different weight loss methods or about different medical treatments for a particular disease because you or someone related to you wants to lose weight or has been diagnosed with a certain medical condition. In both cases, besides consulting a medical or health expert or asking friends or relatives, in today's digital age you would most likely search the World Wide Web (WWW) in order to find information on these issues.

Since the advent of the WWW in the early 1990s, the amount of information available online has grown exponentially. The technical development since then enables individuals worldwide to comfortably and instantly retrieve information on almost any topic and in almost any representation format imaginable (e.g., text, pictures, audio, video). Therefore, in recent years for many people the Web has become a major information resource both in professional and personal life. The vignette presented above about searching information on weight loss methods or medical treatments demonstrates that besides searching for simple and uncontroversial facts or researching product purchases, today the Web also serves as a rich information resource for people conducting research on more complex academic or science-related topics (Horrigan, 2006). In the context of personal or social concerns of individuals, such as medicine and health care – which is the domain of focus in the present dissertation – or environmental, technical or political issues, consulting the Web as a supplement to the interaction with experts has achieved great popularity among laypersons[1], that is, non-experts in a domain (Fox, 2006; Fox & Jones, 2009; Morahan-Martin, 2004; Smith, 2009; Stadtler & Bromme, 2007).

To discover and access information on the Web, most people start by using a general search engine such as Google, Yahoo, or Bing. According to the German ARD/ZDF-Online-Study 2009 (Van Eimeren & Frees, 2009) search engines are the most used applications on the Internet, lying even up with email usage. The latest report from

---

[1] According to Stadtler and Bromme (2007) laypersons are characterized by the fact that they aim at developing a basic understanding of the relevant concepts of a topic of interest, but not at becoming experts in the given field of knowledge.

the PEW Internet & American Life Project (Fallows, 2008) revealed similar results, with almost half (49%) of American Internet users employing search engines on a "typical day" (only exceeded by email usage with 60%).

However, finding relevant, high-quality information on the Web is not an easy task, due to the specific characteristics of the WWW: Although for almost any topic an unprecedented wealth of information can be accessed on the Web, there is no guarantee to its validity and reliability. 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 (Metzger, 2007). Furthermore, contrary to traditional information sources such as printed publications, documents on the Web other than electronic copies of traditional publications seldom have explicit editorial review policies or undergo quality controls (Metzger, 2007). As a result, the quality of online information, for example, on topics like medicine and health care, varies considerably, with many Web sites containing one-sided, biased, or even false information (Bates, Romina, Ahmed, & Hopson, 2006; Eysenbach, Powell, Kuss, & Sa, 2002). Thus, Web users themselves are responsible for "gatekeeping", that is, appropriately evaluating the quality of information found on the Web (Bråten & Strømsø, 2006a; Britt & Aglinskas, 2002) in order to "separate the wheat from the chaff" – in particular when dealing with complex and controversial science-related issues, such as, for instance, the effectiveness of different weight loss methods or medical treatments.

Such a normative perspective concerning the evaluation processes searchers *should* engage in during Web search, however, has to be distinguished from a descriptive perspective concerning the evaluation processes searchers *actually do* display spontaneously during Web search. Yet, the current state of the art provides no conclusive evidence whether or under which preconditions users spontaneously engage in evaluations of information quality when searching for science-related topics on the Web. Accordingly, a major goal of the present dissertation was to investigate what type of spontaneous evaluation processes occur when laypersons search the Web to find information about a complex medical or health-related issue.

The first research question of this dissertation was therefore:

1. What type of spontaneous evaluation processes do laypersons engage in on search engine results pages (SERPs) and on Web pages during Web search for medical and health information?

Furthermore, the dissertation aimed at identifying factors that potentially facilitate the critical evaluation of information quality during Web search. Based on a conceptual framework proposed by Lazonder and Rouet (2008; see also Rouet, 2006) three different types of variables that may influence the evaluation of information quality during Web search were examined, namely (a) contextual variables, (b) individual variables, and (c) resource variables.

More precisely, three further research questions addressed by this dissertation were:

2. Which role do evaluation instructions given before the task (contextual variable) play in laypersons' evaluation of information quality during Web search on a complex medical issue?

3. Which role do user characteristics (individual variables), such as, prior domain knowledge and epistemic beliefs, play in laypersons' evaluation of information quality during Web search on a complex medical issue?

4. Which role does the search interface (resource variable) play in laypersons' evaluation of information quality during Web search on a complex medical issue?

These four research questions were examined in three experimental studies with university students searching for complex medical and health-related issues on the Web. The first study examined laypersons' spontaneous evaluation processes on SERPs and on Web pages and furthermore dealt with the influence of evaluation instructions given before the task (contextual variable) and prior domain knowledge (individual variable) on these evaluation processes. The second and third study, then, were concerned with the effects of personal epistemology (individual variable) and the search interface a Web user has at his or her disposal to access information sources (as a resource variable) on evaluations of information quality on SERPs.

Chapter 2 of this dissertation will outline the characteristics of information search on the Web and the cognitive processes involved, with a focus on complex science-related Web searches. The chapter starts with a definition of information search as a problem solving process, differentiating complex from simple information problems.

Then, characteristics of medical searches on the Web and the issue of information quality will be addressed. Furthermore, the cognitive processes of information search on the Web will be described on the basis of the IPS-I (Information Problem Solving using Internet) model by Brand-Gruwel, Wopereis, and Walraven (2009). The chapter concludes with the differentiation of three different phases in which evaluation processes are assumed to be carried out during information problem solving on the Web. Chapter 3 presents the information foraging theory (Pirolli, 2007; Pirolli & Card, 1999) from cognitive science and the documents model (Britt, Perfetti, Sandak, & Rouet, 1999; Perfetti, Rouet, & Britt, 1999; Rouet, 2006) from text comprehension research as two central theoretical frameworks for evaluation processes during Web search. Furthermore, it reviews empirical findings from these research fields as well as from information science literature and addresses the influence of prior domain knowledge on evaluation processes as well as methodological considerations about the analysis of spontaneous evaluation processes during Web search. Chapter 4 describes Study 1, which empirically investigated spontaneous evaluation processes on SERPs and on Web pages during Web search for a complex health-related issue with a standard search engine. Moreover, Study 1 investigated the influence of evaluation instructions and prior domain knowledge on evaluation processes. Chapter 5 starts with a short summary of the results of Study 1 and then presents theoretical considerations as well as empirical findings concerning the role of personal epistemology and of the search interface in the evaluation of information quality during Web search. Chapters 6 and 7 describe two experiments on the effects of personal epistemology and the search interface on the evaluation of information quality on SERPs during Web search on a complex medical issue. The experiments differ with regard to the resource variable (i.e., the search interface), investigating two different interface design approaches assumed to facilitate the evaluation of information quality. Study 2, reported in Chapter 6, examined the effects of an alternative layout of search results that aimed at reducing the salience of the ranking of the search results. Besides changes in the layout of the search results, the search interface in Study 3, reported in Chapter 7, was augmented by additional quality-related information. This design approach was assumed to guide laypersons in their Web search and to support them in making evaluations of information quality. Chapter 8, the final chapter, summarizes and discusses the findings of the present dissertation. Theoretical and practical

implications of the findings as well as suggestions for further research will be presented.

2   Characteristics and cognitive processes of Web search on science-related information

This chapter outlines the characteristics of information search on the Web and the cognitive processes involved, with a focus on complex science-related Web searches. First, information search will be defined as a problem solving process and complex science-related search tasks will be distinguished from simple fact-finding tasks (section 2.1). Then, characteristics of medical searches on the Web and the issue of information quality will be addressed (section 2.2). Furthermore, in section 2.3 the cognitive processes of information search on the Web will be described on the basis of the IPS-I (Information Problem Solving using Internet) model by Brand-Gruwel, Wopereis, and Walraven (2009). This section concludes with the differentiation of three different phases in which evaluation processes are assumed to be carried out during information problem solving on the Web.

2.1   Information problem solving on the Web

Many researchers consider the process of information search (also known as information seeking) as a problem-solving process driven by an information problem (e.g., Brand-Gruwel et al., 2009; Brown, 1991; Marchionini, 1989, 1995; Kuhlthau, 1993; Wilson, 1999; Wopereis, Brand-Gruwel, & Vermetten, 2008). Thus, information search is a goal-driven activity, which can be distinguished from serendipitous browsing, in which Web users are traversing information in a random and undirected way without having a concrete goal (i.e., an information problem) in mind.

An information problem arises when a discrepancy occurs between information needed to answer a question and information already known (Walraven, Brand-Gruwel, & Boshuizen, 2009). Information problems can be simple, such as in the case of finding a specific fact, or complex, for example, when seeking information to support an argument or to make a complex decision (Marchionini, 1992). Accordingly, simple fact-finding tasks and more complex information search tasks are two broad categories, in which information searches on the Web are usually

classified (e.g., Kellar, Watters, & Shepherd, 2007; Navarro-Prieto, Scaife, & Rogers, 1999; Rose & Levinson, 2004; Shneiderman, 1997).

Fact-finding tasks are characterized as specific, closed-ended tasks (e.g., Tu, Shih, & Tsai, 2008). They include simple lookup of concrete facts about a specific topic such as the height of the Empire State Building, the states of Germany, or the components of a computer system. Thus, such tasks refer to well-structured problems that usually possess one correct answer or a set of convergent answers (Jonassen, 1997). In cognitive science, simple fact-finding tasks are standard paradigms used to study cognitive processes during Web search (e.g., Brumby & Howes, 2008; Miller & Remington, 2004).

In contrast, in complex search tasks – which are in the focus of the present dissertation – the user's goal typically is to form own opinions or to make an informed decision about a controversial, often science-related topic. Complex search tasks address ill-structured problems that do not possess clear-cut, right or wrong solutions. Instead, ill-structured problems are characterized by fragile and conflicting evidence as well as by competing perspectives and arguments, which (all) have to be considered to make an informed decision (Jonassen, 1997). Therefore, complex search tasks are also defined as research-based, open-ended search tasks (e.g., Tu et al., 2008). They require searchers to collect, evaluate, and integrate information from multiple sources (Aula & Russell, 2008). Accordingly, they are usually more cognitively demanding and time-consuming than the aforementioned fact-finding tasks. Psychological research on the cognitive processes involved in Web search for complex science-related issues, however, is still in its infancy.

Typical examples of ill-structured problems and, thus, complex search tasks are medical and health-related issues, for which usually a vast body of conflicting evidence exists and personal decisions under uncertainty have to be made (Eysenbach, 2008, Kienhues, Stadtler, & Bromme, 2011; Stadtler & Bromme, 2007, 2008).

2.2 Web search on medical and health information: A plethora of information of highly variable quality

With the proliferation of medical and health Web sites on the WWW today allowing searchers to quickly and easily access several billions of medical and health Web

pages, Web search for medical and health information constitutes a highly popular activity. According to survey data from the PEW Internet & American Life Project (Fox & Jones, 2009) and the study "Health Care Monitoring 2009" (YouGovPsychonomics AG, 2009) of the German market research institute YouGovPsychonomics AG, in 2009 83% of American Internet users and 79% of German Internet users searched for medical and health information on the Web. However, although being a popular activity, Eysenbach (2008) and also Fox and Jones (2009) stress that medical and health-related searches for most people are not necessarily a frequent, every-day activity like searches for shopping or entertainment issues. Therefore, many people might have insufficient experience and expertise with medical and health Web sites (Eysenbach, 2008). People who use the Web to retrieve medical and health information are predominately laypersons, who usually start their search by using a general search engine such as Google or Yahoo (Fox, 2006; Morahan-Martin, 2004). Approximately half of the information searches they conduct are on behalf of someone else's medical and health situation (Fox & Jones, 2009). Web searches for a "certain medical treatment or procedure" or "diet, nutrition, vitamins, or nutritional supplements" are among the most popular topics (Fox, 2006). Two concrete examples of these search topics – that will also be used in the experimental studies reported in this dissertation – are to search for information about benefits and risks of competing therapies for Bechterew's disease (i.e., a chronic inflammatory rheumatic disease affecting the spine) or about healthy and effective weight loss methods. Entering the search terms "therapies Bechterew's disease" or "healthy, effective weight loss" in Google yields approximately 85,800 and 20,800,000 results, respectively, indicating the abundance of information available on the Web regarding these topics. Furthermore, for both topics various different therapies (e.g., physical therapy, radiation therapy, medical therapy, etc.) or weight loss methods (e.g., low fat diets, low carb (carbohydrate) diets, the cookie diet, etc.) are proposed and controversially discussed on the Web.

As the information retrieved from the Web might strongly influence laypersons' medical decisions and health care choices, for instance, on which therapy or weight loss method to choose or to refuse, the use of misinformation is fraught with risks and dangers (Eysenbach, 2008; Fox, 2006). Thus, when seeking for medical and health-related issues on the Web not only the topical relevance (i.e., whether or not a

piece of information addresses the information problem at hand), but also the quality of the information retrieved from the Web is of great importance.

Tate (2010; see also Alexander & Tate, 1999), who proposed a multidimensional quality framework for the Web, defined information quality by means of the following five dimensions: (1) *authority*, that is, the extent to which the author of the information can be identified as having definitive knowledge of the respective subject area, (2) *accuracy*, that is, the extent to which information is reliable and free of errors, (3) *objectivity*, that is, the extent to which information is presented without distortion by personal feelings or other biases (e.g., commercial interests), (4) *currency*, that is, the extent to which information can be identified as up-to-date, and (5) *coverage of information and intended audience*, that is, the breadth and depth of the information and the target audience for whom the information was created. A concept that is closely related to information quality is *credibility* which is composed of *expertise* and *trustworthiness* (cf. Danielson, 2006). Expertise refers to the perceived ability of the information source to provide accurate information, thus relating to Tate's (2010) criteria of authority and accuracy. Trustworthiness refers to the perceived willingness of an information source to provide accurate and unbiased information. Thus, it relates to Tate's criteria of accuracy and objectivity. Hence, credibility can also be defined as a central component of information quality (e.g., Rieh & Danielson, 2007; Hilligoss & Rieh, 2008). Figure 1 represents this perspective on the multidimensional construct of information quality that integrates Tate's dimensions of information quality and the concept of credibility (Danielson, 2005; Fogg et al., 2001).
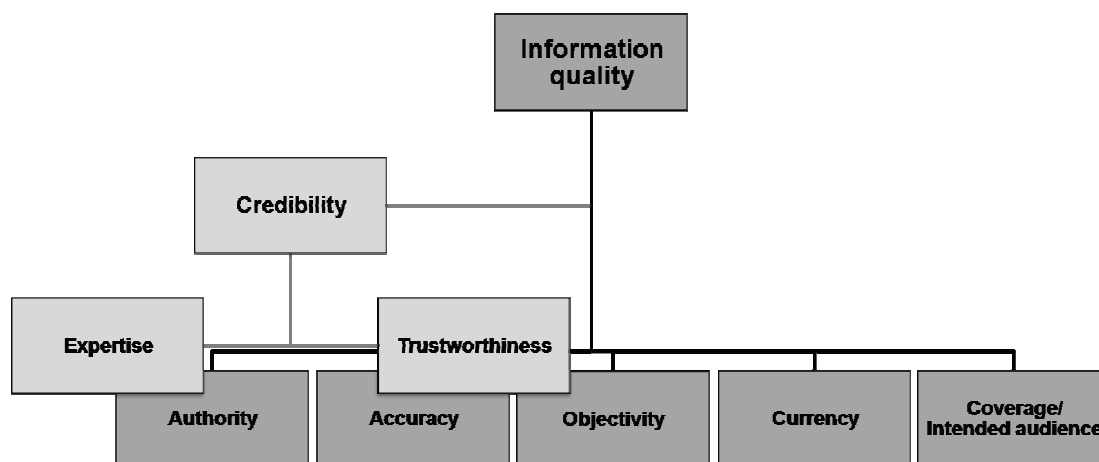
Figure 1. The multidimensional construct of information quality.

In the last decade, various empirical studies have shown that the quality of medical and health information on the Web varies considerably, with many Web sites containing one-sided, biased, outdated, or even false information (e.g., Bates, Romina, Ahmed, & Hopson, 2006; Eysenbach, Powell, Kuss, & Sa, 2002). Reasons for the high variability of information quality on the Web are the lack of quality controls (e.g., editorial boards, peer review) coupled with the low financial and technical requirements to publish information online (Metzger, 2007). Accordingly, information providers on the Web differ dramatically in their levels of expertise as well as in their perspectives and motives (Eysenbach, 2008; Metzger, 2007). Besides scientific and other institutions, on the Web also journalists, companies, and laypeople provide medical and health information. Despite the high variability of information quality, the results lists returned by search engines usually present different types of information sources interspersed with each other (e.g., scientific and other official articles, discussions by laypeople, or commercial reports by companies). Thus, even the information contained in Web pages listed among the top search results of a search engine results page (SERP) might turn out to be subjective, unscientific, one-sided, or commercially biased rather than scientifically sound and objective. Mansell and Read (2009), for instance, could show that 42% of the top 54 Web sites about posttraumatic stress disorder returned by Google or Yahoo were funded by drug companies. These Web sites gave significantly more emphasis to medication in the treatment of posttraumatic stress disorder than the rest of the Web sites. Furthermore, Hwang et al. (2007) investigated the quality of weight loss advice exchanged in 18 Internet forums that were among the top 50 search results returned by Google. Results showed that 23% of the advice given in the forums was erroneous or at least not in line with scientific weight loss guidelines. Misinformation was most likely included in medication-related advice or in advice that was posted in low-activity forums.

To conclude, filtering out relevant and high-quality information, that is, "separating the wheat from the chaff" is crucial when searching for medical and health information on the Web. In other words, the critical evaluation of the quality of information sources or the information therein, also referred to as source evaluation in the literature (Bråten, Britt, Strømsø, & Rouet, 2011; Harris, 2010; Wiley et al., 2009), constitutes a central component of (medical) information problem solving on

the Web. Otherwise searchers might base important personal decisions on one-sided, biased, or even false information.

In the remainder of this dissertation the terms *evaluation of information quality* and *source evaluation* will be used interchangeably, contrasted to the *evaluation of topical relevance*, that is, whether or not a piece of information addresses the topic at hand. The term *evaluation processes* refers to both types of evaluation brought to bear during Web search.

The following section presents a model that describes the process of how a competent Web searcher solves information problems on the Web. Furthermore, three different phases will be identified in which evaluation processes are assumed to be carried out during information problem solving on the Web in order to judge the perceived topical relevance and quality of information.

2.3   The process of information problem solving on the Web

Marchionini (1989) has pointed out that "Information-seeking is a special case of problem solving, [that …] includes recognizing and interpreting the information problem, establishing a plan of search, conducting the search, evaluating the results, and if necessary, iterating through the process again" (p. 54). Following Marchionini (1989), during the last two decades several models in educational psychology and human-computer-interaction have described the information problem-solving process by segmenting it into several sub-processes. Whereas the majority of the models like Marchionini (1989) addressed information seeking in general (e.g., Kuhlthau, 1991; 1993; Sutcliffe & Ennis, 1998; Wilson, 1999), some researchers have tried to develop models specifically tailored for Web search (e.g., Broder, 2002; Hölscher & Strube, 2000; Marchionini & White, 2008; Shneiderman, Byrd, & Croft, 1997, 1998).

One of the most recent and most comprehensive of these models is the IPS-I model proposed by Brand-Gruwel et al. (2009). The IPS-I-model describes the process of information problem solving (IPS) on the Internet (I) assumed to be performed by a competent Web searcher. The model proposes five main steps that are typically performed in an iterative fashion: (1) define an information problem, (2) search information, (3) scan information, (4) process information, and (5) organize and present information; see Figure 2.
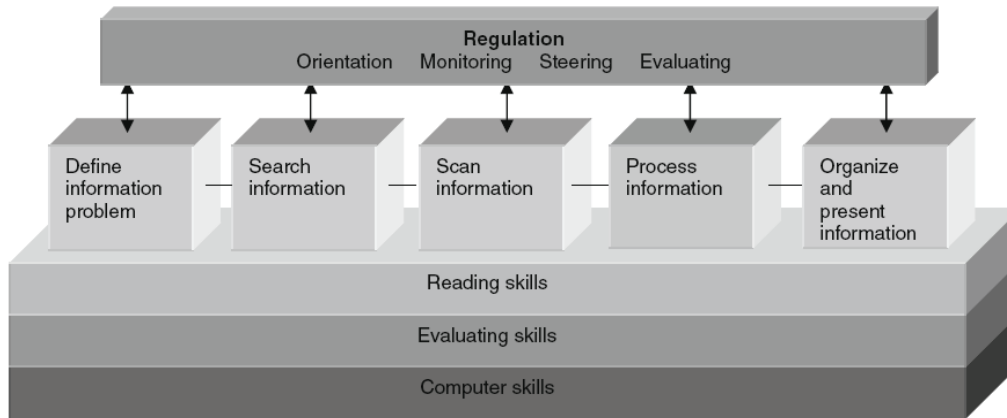
Figure 2. The information problem solving using internet model (IPS-I-model) by Brand-Gruwel et al. (2009).

The process of information problem solving starts with the recognition of an information need, which is transformed into a concrete and comprehensive problem definition ('define information problem'). An example would be to decide to search for information about how to lose weight in a healthy and effective way. While defining the problem, a competent searcher is assumed to activate prior knowledge on the subject matter, that is, in the case of weight loss methods what he or she knows about diets and nutrition. The step 'search information' consists in selecting an appropriate search strategy, that is, most likely to use a search engine, and to specify some search terms, which then have to be entered into the search engine. For the weight loss method example, possible search terms are for instance "weight loss", "healthy", and effective". As a response to the search terms entered, the search engine returns a rank-ordered list of search results with the most relevant and most popular documents typically being the highest-ranked ones (cf. Cho & Roy, 2004). On the basis of the search result descriptions comprising a title, an excerpt of the content of the Web page, and its URL (uniform resource locator) the search results are assumed to be judged with regard to both topical relevance and quality, in order to select those search results for further inspection that seem useful to solve the information problem. Figure 3 is an example of a SERP for the search terms "weight loss healthy effective" entered to Google (retrieved July 15, 2010).

Figure 3. Search engine result page for the search terms "weight loss healthy effective".

The first search result presented on the SERP in Figure 3 indicates that the corresponding Web site addresses natural and holistic approaches to lose weight. The URL www.womentowomen.com suggests that on the Web site women exchange their personal experiences with regard to weight loss methods. The fifth search result, in contrast, links to the "Weight Watchers" Web site, a well known commercial organization, which can be assumed to promote their Weight Watchers Diet. Then, further down the list there are several search results which can be assumed to link to official Web sites providing scientific articles or study results,

13

such as the "Science Daily" article (www.sciencedaily.com) about a new effective weight loss revealed according to a new study.

After accessing a selected Web page, the step 'scan information' proposes that the searcher first scans and evaluates the Web page in order to get an idea of the kind of information provided, and in order to decide whether it is useful. If the Web page is not perceived to be useful, the searcher is assumed to return to the SERP and to select another search result. Alternatively, the searcher might refine the search by entering new or additional search terms. For example, in the case of the first search result, the searcher might find out that "Women to Women" stands for a health care center providing various health care services that are subject to fees, such as a personal diet program (see Figure 4). Therefore, one might conclude that the information providers have commercial interests in promoting their weight loss program, and as a consequence to reject the further processing of this Web site.
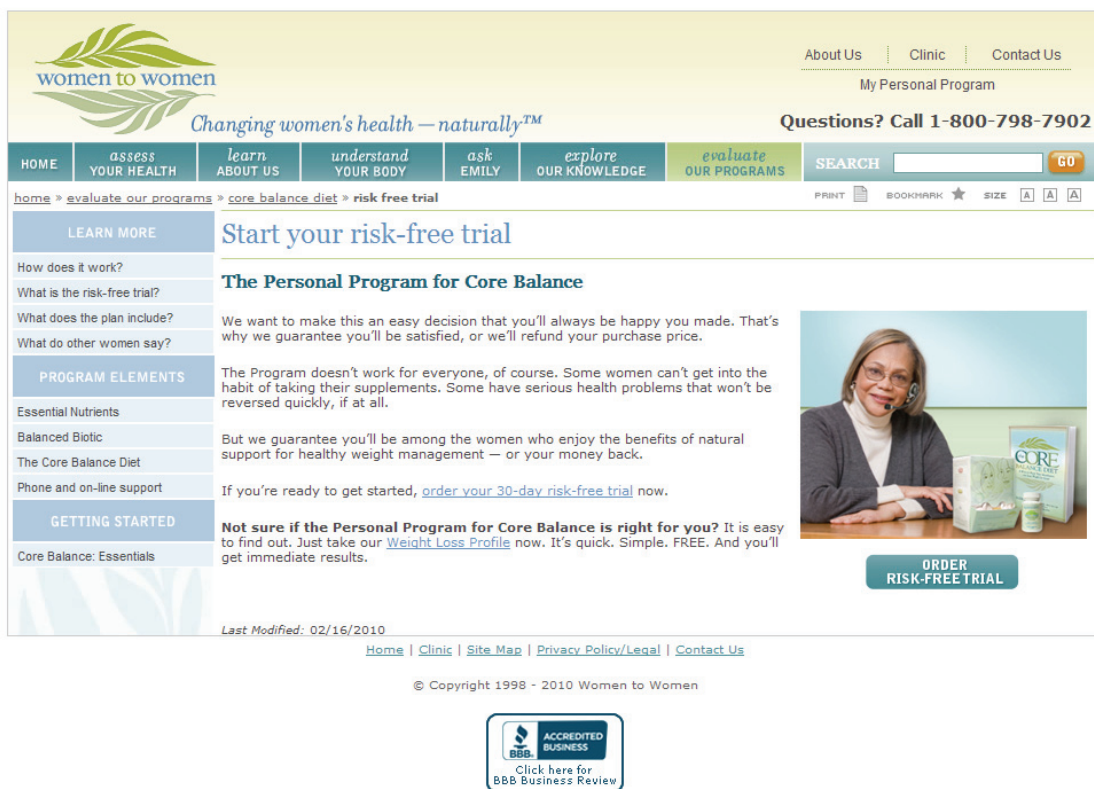
Figure 4. Personal diet program offered on the "Women to Women" Web site.

In case that a Web page is judged as being useful, the step 'process information' consists in structuring and elaborating on the content, integrating the found information with prior knowledge, and in evaluating the processed information with regard to quality and coherence. Typically, in particular in the case of complex ill-

structured information problems, further Web pages are assumed to be accessed, scanned and processed. Whereas the first four steps of the model are part of an analysis phase, the fifth step 'organize and present information' concerns the synthesis of information. In particular when solving complex ill-structured information problems, this step involves the comparison of information found across various documents, to resolve conflicts and incoherencies within the documents collection, and to integrate the information towards a solution of the information problem, for instance, to decide which weight loss method one would start to undergo. The solution can be represented internally in the user's mind or presented externally through oral or written communication.

In addition to the five processing steps, regulation activities (represented by arrows in Figure 2) are assumed to be carried out during the entire IPS process. These regulation activities such as orienting on the task, monitoring and steering performance, time management, and assessing the process as well as the final solution play a key role when it comes to the efficiency and effectiveness of the Web search process and its outcome. To appropriately perform the five main skills as well as the regulation activities, Brand-Gruwel et al. (2009) further propose that Web searchers have to possess some basic prerequisite skills, namely, (a) hypertext reading skills, (b) computer skills, and (c) evaluating skills (see the three bottom layers in Figure 2). Besides such *individual variables*, that is, searchers' cognitive and metacognitive prerequisites like general reading and comprehension skills, prior domain knowledge, or Web search experience, Lazonder and Rouet (2008) propose two more types of variables that may influence activities during information problem solving and thus should be represented in cognitive models of IPS. These variables are *contextual variables* such as the task type, the problem statement, certain instructions given, or time constraints, and *resource variables* such as the amount and type of information available during the search and the interface or tools available to access information sources (see Figure 5).
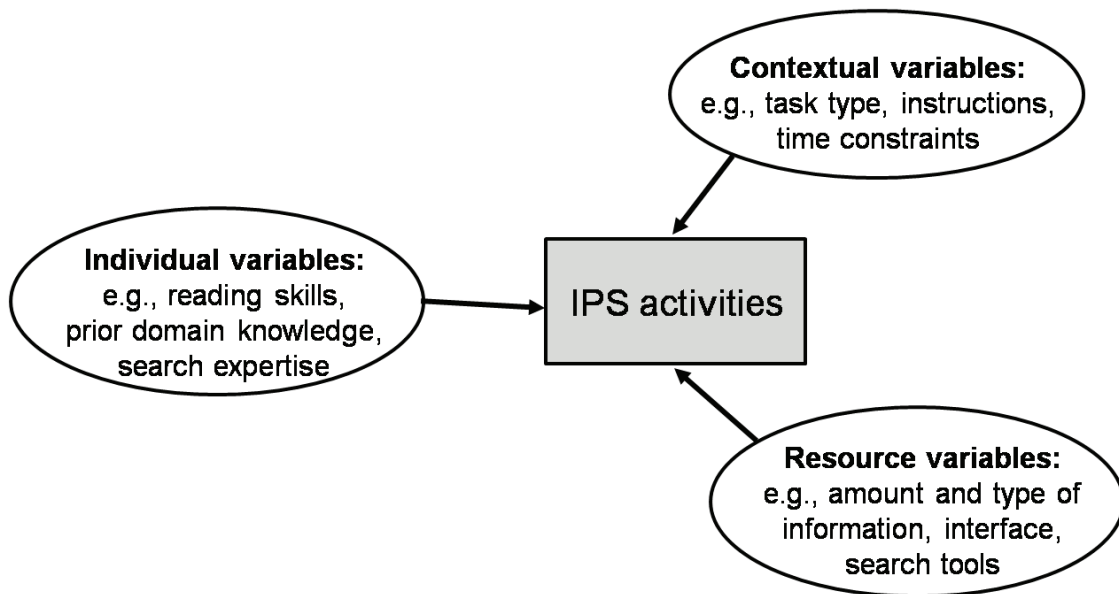
Figure 5. Three types of variables that may influence activities during information problem solving. Figure based on Lazonder and Rouet (2008)

Lazonder and Rouet (2008) consider the evaluation of information as "the heart of the whole IPS activity" (p. 758). During the whole IPS process three phases of evaluation can be distinguished. Initially, Rieh (2002) differentiated between (1) the evaluation of search results provided by the search engine and (2) the evaluation of Web pages (see also Crystal & Greenberg, 2006). Both of these phases are assumed to involve evaluations of topical relevance and information quality. Because the evaluation of search results is based on only sparse information about the corresponding Web pages and the information they contain, according to Rieh (2002) in this evaluation phase *predictive judgments*, that is, predictions about the usefulness of available documents for the search task at hand, are made. When having accessed a Web page, according to Rieh (2002) searchers make *evaluative judgments* about the Web page and the information therein. When the evaluative judgment matches their expectations made in the predictive judgment, searchers process the information on the Web page in detail; otherwise they leave the Web page. Recently, Rieh and Hilligoss (2008) have extended this framework of predictive and evaluative judgments, by adding a third evaluation phase called *verification*. Verification refers to the re-evaluation of information when searchers encounter contradictory information in a collection of information sources. Thus, this third evaluation phase can be assumed to mainly address the evaluation of information quality.

To sum up, during Web search on complex issues three pivotal evaluation phases that unfold in an iterative manner can be considered: (1) the evaluation of search results, (2) the evaluation of Web pages, and (3) the evaluation of document collections when comparing and integrating multiple documents. The remainder of this dissertation will focus on the theoretical and empirical examination of the evaluation processes regarding topical relevance and information quality brought to bear during these evaluation phases.

Two central theories which address the type of evaluation processes in question are the *information foraging theory* by Pirolli and Card (1999; see also Pirolli, 2007) from cognitive science and the *theory of documents representation* (i.e., the documents model framework) by Perfetti et al. (1999; see also Britt et al., 1999; Rouet, 2006) from text comprehension research. Whereas the information foraging theory in particular addresses the first two evaluation phases of selecting search results and visiting Web pages, the theory of documents representation is especially related to the latter two evaluation phases of evaluating information within documents (e.g., a Web page) and across multiple documents (e.g., several Web pages). Both theories will be outlined in the next chapter.

3 Evaluation processes during Web search and the role of prior domain knowledge and of evaluation instructions

Chapter 3 presents the information foraging theory (Pirolli, 2007; Pirolli & Card, 1999) from cognitive science (section 3.1) and the theory of documents representation (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006) from text comprehension research (section 3.2) as two central theoretical frameworks for evaluation processes during Web search. Furthermore, evaluations of topical relevance and of information quality will be contrasted and the impact of prior domain knowledge as an important individual variable on evaluations of information quality will be addressed. Moreover, the chapter outlines empirical findings from information science literature on evaluation processes on the Web (section 3.3) and addresses methodological considerations with regard to the effects of the instructions given for a search task (contextual variable) on evaluation processes during Web search (section 3.4). The theoretical, empirical, and methodological elaborations provided will be used to derive the research questions of Study 1 of this dissertation.

3.1 Information foraging theory: Search result selection based on information scent and satisficing strategies

In cognitive science, one of the most influential theories describing the cognitive processes involved in hyperlink selection (e.g., the selection of search results) in order to explain and predict Web search and navigation behavior is the information foraging theory (Pirolli, 2007, Pirolli & Card, 1999). According to Pirolli and Card (1995) information foraging refers to activities associated with evaluating, seeking, and handling information sources.

One of the core concepts of information foraging theory is the concept of *information scent*. This concept is based on Brunswik's lens model (1956; as cited in Pirolli, 2007), which postulates that the judgment of distal sources is indirect and has to be based on proximal cues. Accordingly, during Web search based on the proximal cues available in search results (or hyperlinks), Web users have to make a predictive

judgment (cf. Rieh, 2002) about which search result may lead to a desired distal source (i.e., a Web page). Information foraging theory postulates that this judgment is determined by the strength of the information scent. Information scent reflects the perceived semantic similarity between the proximal cues (i.e., keywords or trigger words) available in the search results and the current search goal of the user, which is defined by a desired information that is expected to be located in a distal information source (e.g., a Web page). In other words information scent is the perceived semantic (i.e., topical) relevance of screen objects to a user's current information need (e.g., Fu & Pirolli, 2007; Juvina & Van Oostendorp, 2008). A strong information scent of a search result indicates a high likelihood that the respective distal source contains the desired information and thus increases the likelihood that the search result will be selected.

According to information foraging theory, a Web user's evaluation of information scent is based on spreading activation (cf. Anderson, 1983) in a semantic memory network that represents the Web user's declarative knowledge. A strong information scent occurs when the encoding of proximal cues in semantic memory results in a substantial spread of activation to the representation of the current search goal. Figure 6 illustrates the concept of information scent (IS) for a user pursuing the goal of finding information about "medical treatments for Bechterew's disease" (this is the desired distal information defining the search goal). It is assumed that the user encounters a search result like the one depicted in Figure 6, which includes terms like "patients, inflammation, spine, physical therapy" (these are the available proximal cues).

Figure 6. Example of a search result containing the proximal cues patients, inflammation, spine, and physical therapy.

The arrows in Figure 7 represent the spread of activation from these terms in the search result to the goal representation, which is used to calculate the information scent of the search result. More precisely, each proximal cue in the search result spreads activation to the goal representation in proportion to its strength of

association with the goal representation. For instance, the term medical (in the goal representation) and the term patient (in the search result) have a high strength of association in semantic memory, because the two terms co-occur frequently in the environment.
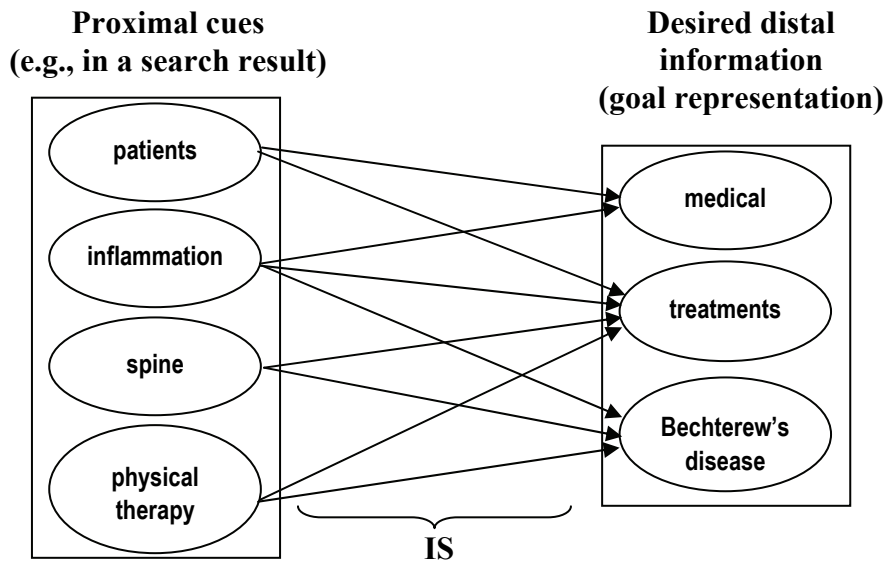
Figure 7. Illustration of information scent (IS), example adapted from Pirolli (2007).

Furthermore, information foraging theory assumes that Web searchers aim at maximizing gains of valuable information in relation to their effort invested (i.e., search time, cognitive effort). This leads to a *satisficing strategy*, a form of bounded rationality (cf. Simon, 1955), that implies that users do not evaluate the information scent of all search results available, but evaluate search results only until one is encountered that is "good enough". In line with this assumption, several descriptive studies have shown by means of eye tracking and/or log file analyses that search engine users (required to solve simple fact-finding tasks) tend to read search results in the order in which they were presented by the search engine (Cutrell & Guan, 2007; Granka, Joachims, & Gay, 2004; Joachims, Granka, Pan, Hembrooke, & Gay, 2005), with the second result page being only rarely visited (e.g., Lorigo et al., 2006). Moreover, searchers spend most attention to the search results on top of a SERP and also predominantly select these first few links (Cutrell & Guan, 2007; Granka et al., 2004; Guan & Cutrell, 2007; Joachims et al., 2005; Pan et al., 2007).

Once a search result is selected and thus an information source is accessed, this source is 'consumed' by processing it until the information scent stops getting stronger. That is, information processing is terminated when the user no longer expects to find additional useful information in the information source. In this case, the searcher either moves on to a different information source with a higher information scent or ends the search (Pirolli, 2007).

To sum up, information foraging theory (Pirolli, 2007; Pirolli & Card, 1999) explains Web search and navigation behavior (i.e., link selections and site leaving actions) based on the notions of information scent and satisficing strategies. Thus, the theory presupposes that Web searching is guided by the perceived topical relevance of Web information and respective cost-benefit analyses.

Based on information foraging theory and thus on the concepts of information scent and satisficing strategies (incorporating link position), Fu and Pirolli (2007) have developed a computational cognitive model called SNIF-ACT (Scent-based Navigation and Information Foraging in the ACT cognitive architecture) to predict users' Web search and navigation behavior. Monte Carlo simulations of the SNIF-ACT model showed good fits to actual user data collected in a controlled study on realistic Web search tasks. Similarly, other computational models of Web navigation based on the concept of information scent (e.g., CoLiDeS by Kitajima, Blackmon, & Polson, 2000; CoLiDes+ Pic, Karanamn, Van Oostendorp, Puerta Melguizo, & Indurkhya, 2009, that also considers semantic information from pictures), a combination of information scent and navigation strategies (e.g., CoLiDeS+ by Juvina & Van Oostendorp, 2008), or a combination of information scent and satisficing mechanisms (e.g., MESA by Miller & Remington, 2004) also have been able to successfully predict Web search and navigation behavior in several studies.

The evaluation of information quality is ignored completely by information foraging theory and respective computational models. However, it can be claimed that the tasks used for modeling in the aforementioned studies were designed in a way that they required users to focus their attention on the topical relevance of available information: Users either had to engage in simple fact-finding tasks or they had at their disposal a selection of Web information that was restricted to uncontroversial and consistent information of established quality. It seems plausible, that under these conditions, the evaluation of information quality is not a precondition for a model to

successfully predict search behavior, or for a user to successfully perform the task. Yet, as has been outlined in Chapter 2 it can be assumed that the evaluation of information quality needs to be considered (1) when the search task is sufficiently complex and, even more important, (2) when the available information is highly variable with regard to its quality, as it is usually the case when searching for complex science-related issues, such as medical or health-related issues, on the Web (cf. Chapter 2.2). Thus, information foraging theory might be insufficient to explain users' Web search and evaluation behavior when accomplishing complex Web search tasks about science-related issues for which information of variable quality is encountered on the Web (cf. Gerjets & Kammerer, 2010). A theoretical framework that addresses the evaluation and use of diverse sources of information when accomplishing complex tasks about science-related issues is the *documents model framework* proposed by Perfetti and colleagues in their theory of documents representation (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006). The documents model will be outlined in the following section.

3.2   The documents model framework: Source evaluations during multiple documents reading

With the "documents model" Perfetti and colleagues (1999) proposed a cognitive model to describe how competent readers (Rouet, 2006) process multiple, diverse documents to learn about a complex topic at hand. Although the documents model framework originated from studies about history learning with printed documents (e.g., Britt & Aglinskas, 2002; Nokes, Dole, & Hacker, 2007; Rouet, Britt, Mason, & Perfetti, 1996; Wineburg, 1991), in the last decade it has been successfully applied to multiple-text reading tasks in several other domains such as medicine and health care (Sanchez et al, 2006; Stadtler & Bromme, 2007, 2008; Wiley et al., 2009), climatology (Bråten, Strømsø, & Britt, 2009; Strømsø, Bråten, & Britt, 2010), and psychology (Le Bigot & Rouet, 2007). Furthermore, because during Web search for complex science-related issues information has to be retrieved from multiple sources that might express diverse or even contradictory viewpoints (cf. Aula & Russel, 2008; Brand-Gruwel et al., 2009) the documents model framework has been recently applied to reading multiple Web-based documents as well (e.g., Le Bigot & Rouet,

2007; Salmerón, Gil, Bråten, & Strømsø, 2010; Sanchez et al., 2006; Stadtler & Bromme, 2007, 2008; Wiley et al., 2009).

3.2.1 Representing and processing multiple documents

The documents model can be regarded as an extension of Kintsch's (1988, 1998) construction-integration model of single-text comprehension, which distinguishes between a textbase and a situation model as two central layers of mental representation being constructed during the comprehension of a single text. Whereas the text base contains the meaning of the text itself without adding any additional information, the situation model refers to the interpretation of the text based on inferences drawn from the text by integrating the textbase with the reader's prior knowledge.

According to the documents model, when reading multiple documents, two additional layers of representation are constructed: the *situations model* and the *intertext model*. In the situations model the integrated content from various documents retrieved during information seeking is represented by drawing inferences across texts. Thus, the situations model reflects the overall understanding of the topic at hand including both content uniquely presented in a single text and agreed upon or conflicting content from various texts. In addition, the intertext model contains information about how the various information sources relate to each other (i.e., whether the documents agree, complement, or contradict each other). Furthermore, the intertext model also represents information about the sources in the form of document nodes for each document. In these document nodes information about the document type (e.g., scientific article, private blog), the date and the publisher of the document, the expertise of the document's author as well as the suspected motives of the author (e.g., to inform, to persuade, or to exchange experiences), and the intended audience for whom a document was created (e.g., for experts, lay people, or clients) can be represented. Thus, information represented in the documents nodes address the quality of the document or the information therein (cf. Chapter 2.2). Finally, in a complete documents model, source information in the intertext model is assumed to be linked to the information in the situations model. This allows the reader to interpret information in the light of the source characteristics.

Whereas the original documents model framework proposed by Perfetti et al. (1999) focused on describing the representations of the intertext model and the situations model, Britt and Rouet (2011, as cited in Bråten, Britt, et al., 2011) recently revised the documents model framework by adding assumptions on the underlying strategic processes required by the reader to create these representations. To specify these strategic processes during documents model construction, they elaborated on the so-called corroboration and sourcing heuristics identified by Wineburg (1991) in domain experts.

Corroboration is the process of systematically comparing contents across documents (or to prior knowledge) and thereby identifying consistencies or discrepancies among them (content-content links, see Figure 8). Thus, corroboration helps to create a situations model including agreed upon or conflicting information from various documents. In the example of searching for information about the effectiveness of different weight loss methods on the Web, one document might stress the success of low carb diets and how healthy low carb meals are, whereas another document might report on recent study results revealing that low carb diets are ineffective in the long term and increase the risk of high cholesterol. In turn, the detection of such discrepancies across documents may facilitate the attention to source information (cf. Braasch et al., 2010; Bråten, Britt, et al., 2011; Rouet, Britt, Caroux, Nivet, & Le Bigot, 2009). Thus, corroboration can also affect the construction of an intertext model, and thus the process of sourcing.

Sourcing is the process of identifying and evaluating the source of a document (i.e., creating document nodes) prior to reading its content or after having detected discrepancies across documents. In the weight loss methods example, the information about the success of low carb diets might be identified as being provided by a nutrition shop, whereas the information about the ineffectiveness of low carb diets might be identified as being released by an official institution (e.g., the National Institute of Health). Moreover, sourcing allows to create relationships between information sources (i.e., source-source links, see Figure 8), indicating, for instance, that the National Institute of Health and the nutrition shop disagree on the effectiveness of low carb diets. Finally, sourcing allows to link source information in the intertext model to contents in the situations model (i.e., source-content links, see Figure 8), in order to weight and interpret a document's content in light of its source characteristics. For example, the information provided by the nutrition shop about the

effectiveness of low carb diets for losing weight may be assumed to be biased due to vested interests of the nutrition shop in promoting the treatment. In contrast, the information provided by the National Institute of Health Web site reporting on study results that indicate the ineffectiveness of low carb diets for losing weight might be evaluated as more trustworthy. Hence, the former information should play an inferior role in a competent reader's decision regarding the weight loss method. In other words, on the basis of these two documents a competent reader would likely come to the conclusion that low carb diets are not effective to lose weight and, moreover, might endanger human health. However, in a real situation several other documents providing additional perspectives can be assumed to be taken into account in making the decision at hand.
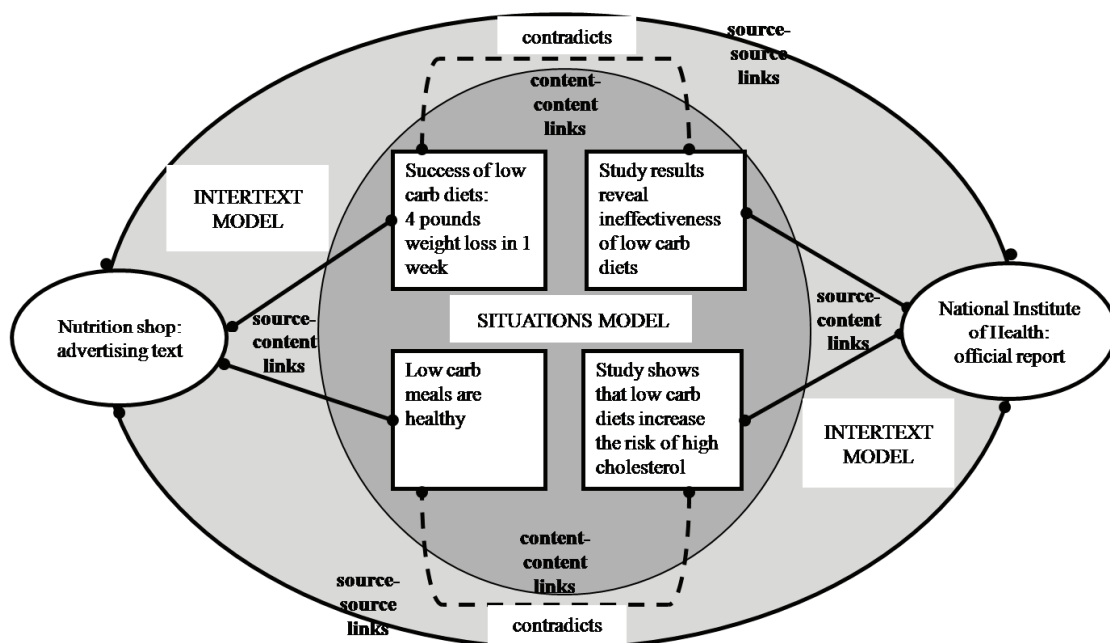
Figure 8. Documents model of two documents about low carb diets. The ovals represent the source information and the rectangles the content of the two documents. The dashed connection lines represent the process of corroboration (with the formation of content-content-links). The solid connection lines represent the process of sourcing (with the formation of source-source links and source-content-links). Figure based on Bråten, Britt, et al. (2011)

Another recent extension of the documents model framework is the MD-TRACE (Multiple Documents – Task-based Relevance Assessment and Content Extraction) model proposed by Rouet and Britt (2011). Similar to the IPS-I model (Brand-Gruwel et al., 2009; see Chapter 2.3) the MD-TRACE model describes the process of

multiple-text comprehension by means of five major processing steps that unfold in an iterative way. In step 1 the goals of the task, that is, the expected outcomes of the reading activity, are defined. Then, in step 2 the reader is assumed to assess his or her information needs by comparing the desired task outcomes with his or her prior knowledge about the problem at hand. Step 3 is a complex step that is considered to involve the evaluation and selection of documents based on their topical relevance and information quality, scanning as well as in-depth content processing, and the comparison and integration of information across documents. Step 4 consists in the creation of a task product (e.g., an answer to a question or an essay), which in step 5 is assumed to be finally evaluated by the reader. Furthermore, the MD-TRACE model considers external resources and internal resources that might affect the processing steps. External resources basically are what Lazonder and Rouet (2008) defined as (1) contextual variables, namely the task specifications, that is, the task itself as well as any instructions given to solve the task, and as (2) resource variables, that is, the set of documents and access devices available, as well as source information provided in the documents. Internal, or cognitive, resources are individual variables (cf. Lazonder & Rouet, 2008) such as general reading, comprehension, and memory skills, prior task experience or prior domain knowledge. The role of prior domain knowledge in source evaluations during multiple-text comprehension will be addressed in detail in the following section.

3.2.2 The role of prior domain knowledge in source evaluations during multiple-text comprehension

Perfetti et al. (1999) and Rouet (2006) postulate that the construction of an integrated and elaborated documents model is strongly influenced by the readers' prior domain (or discipline) knowledge. In line with this theoretical assumption, there is ample empirical evidence that prior domain knowledge affects students' reading of multiple documents, with high prior knowledge leading to better text comprehension and information integration across documents (e.g., Bråten & Strømsø, 2006b; Bråten & Strømsø, 2010a; Gil, Bråten, Vidal-Abarca & Strømsø, 2010; Le Bigot & Rouet, 2007; Pieschl, Stahl, & Bromme, 2008; Rouet, Favart, Britt, & Perfetti, 1997; Strømsø, Bråten, & Samuelstuen, 2008; Wineburg, 1991). With regard to source evaluations, as it will be outlined in the following, prior domain knowledge also has been shown to play an important role.

In his seminal study, Wineburg (1991) investigated expert historians and high-school students reading multiple documents (i.e., eight written texts plus three pictures) about a particular historical event. Thinking-aloud data revealed that whereas expert historians constantly evaluated the source of each document (e.g., regarding the author or document type) prior to reading the contents and actively used such information in their interpretation of the document's content, domain novices often tended to ignore source information.

In another much-cited study Britt and Aglinskas (2002) showed that both high-school and college students' spontaneous attention to source information (e.g., author, date, document type) and their evaluation of sources in terms of trustworthiness or potential bias while reading multiple documents (six texts) about a historical controversy was rather low. Even though college students' sourcing skills (i.e., noting and evaluating the source of a document or the source of information) were at least somewhat higher than high school students' sourcing skills, their source evaluations still were far from ideal. Although not explicitly reported in this study, it can be assumed that both groups of students possessed low prior knowledge on the historical controversy.

Bråten, Strømsø, and Salmerón (2011) examined undergraduate students' trustworthiness evaluations when working with multiple documents (seven texts) on the controversial topic of climate change. Instead of comparing domain novices and domain experts they investigated differences between novices varying in their level of prior domain knowledge (i.e., rather knowledgeable and rather unknowledgeable novices). After having read seven texts about the topic of climate change, participants were asked to evaluate the trustworthiness of each of the texts and to rate the importance of six different evaluation criteria (author, publisher, document type, content, own opinion about the topic, date of publishing) for their trustworthiness evaluations by means of rating scales. For this task, participants were presented with short descriptions of each text, including source relevant information. At this point, it should be noted, that the testing procedure used in this study might have resulted in source evaluations not reflecting what participants would spontaneously do during normal reading (for further details see the discussion in section 3.4).Yet, results of the study indicated that the low-knowledge readers were more likely to trust a rather untrustworthy, potentially biased source (i.e., a company presentation of an oil company) than the high-knowledge readers. Furthermore, whereas the low-

knowledge readers trusted a popular science text from a research magazine, a newspaper article, and the company presentation from the oil company to the same extent, the high-knowledge readers rated the company presentation to be significantly less trustworthy than the other two sources. Moreover, low-knowledge readers reported to put more emphasis on the date of publication for the evaluation of the documents' trustworthiness than high-knowledge readers. In this study, date of publication, however, was a rather irrelevant criterion to evaluate the documents' trustworthiness because there was very little variation in date of publication among the documents, with all documents having been recently published.

To sum up, readers with low prior knowledge on the topic at hand seem to spontaneously engage in source evaluations only to a very low extent. Furthermore, when explicitly asked to evaluate the trustworthiness of documents they seem to have problems, for instance, to evaluate potential biases and tend to base their source evaluations on rather irrelevant or superficial criteria. According to Bråten, Strømsø, and Salmerón (2011) a possible explanation for the poor source evaluations observed for low-knowledge readers might be that readers with low domain knowledge have to invest more effort in comprehending the content of an information source than readers with higher domain knowledge. As a consequence, low-knowledge readers might have less cognitive resources available to engage in profound source evaluations. If this explanation holds to be true, the problem might turn out to be even more severe in a Web-based reading context that adds onto readers the burden of additional cognitive processes such as handling huge amounts of information and keeping track of the navigation path (cf. Dillon, 2002). Furthermore, on the Web, source characteristics such as author or publisher information are often not displayed in a salient way or even are missing completely (cf. Tate, 2010).

In the field of persuasion research, Stanford, Tauber, Fogg, and Marable (2002) and Fogg et al. (2003) examined which criteria and features domain experts and novices reportedly used to evaluate the credibility of health and finance Web sites. Study participants were asked to rate the credibility of pairs of Web sites and to provide written comments about the features of a Web site used to evaluate its credibility. The survey data corroborate the findings by Bråten, Strømsø, and Salmerón (2011) that users with low prior knowledge tend to base their source evaluations on superficial criteria. Whereas domain novices based their credibility evaluations mainly on the Web page design (e.g., colors, layout, pictures), domain experts most

often relied on author or publisher information, followed by credibility evaluations relating to references provided on the sites or based on perceived motives or biases.

However, rather than being provided with a small set of preselected documents that can be read entirely – which corresponds to the scenarios investigated in the abovementioned studies from text comprehension research and persuasion research – during Web search users are often confronted with thousands or even millions of search results returned by a search engine, making it obviously impossible to access and process all of the suggested Web pages. Accordingly, users themselves are responsible for selecting a manageable subset of the potentially most useful information sources for further exploration (cf. Braasch et al., 2009). Therefore, when applying the documents model framework on the Web context, source evaluations have not only to be considered during the evaluation of Web pages, but also in an earlier stage of Web search, namely during the evaluation of search results (cf. Rouet & Britt, 2011). However, source evaluations on SERPs may also heavily depend on searchers' prior domain knowledge as indicated by two studies outlined in the following.

In a case study by MaKinster, Beghetto, and Plucker (2002) retrospective interviews based on screen recordings of the Web search process revealed that undergraduate students with rather high domain knowledge regarding the search task (i.e., Newton's Third Law) engaged in thorough evaluations of the search results. By intensively evaluating the title, the page excerpt, and the URL of a search result high-knowledge students seemed to aim at identifying whether the document was relevant and whether it originated from a reputable source. In contrast, searchers with low prior domain knowledge mostly relied on the rank position as a superficial cue, simply selecting the search results in the order in that they were presented by the search engine.

A recent Web search experiment by Salmerón, Kammerer, LLorens, and García-Carríon (2010) using a procedure that did not rely on students' self-reports showed similar results. In this experiment undergraduate students who were provided with a Google SERP with ten search results on the topic of climate change were instructed to explore the Web pages accessible through the SERP and to finally select the two most suitable pages that could serve as basis for a report on the topic. Results showed that high domain knowledge students selected a relevant and trustworthy Web page

provided by the United Nations significantly more often than low domain knowledge students, even though the Web page's rank was only the third one in the Google SERP.

To conclude, there is first empirical evidence from studies comparing less knowledgeable and more knowledgeable novices with regard to their source evaluation strategies during Web search for complex science-related issues, that making (sophisticated) evaluations of information quality on SERPs and on Web pages depends on searchers' prior knowledge on the search domain or search topic at hand. In other words, at least when a certain amount of prior domain knowledge is given, searchers seem to attend and evaluate source information on search results and on Web pages as proposed by the documents model.

Further evidence for searchers' engagement in source evaluations on SERPs and on Web pages is provided by information science literature outlined in the next section.

3.3   Information science studies: Relevance judgments on SERPs and Web pages

In the field of information science, the evaluation of search results and Web pages is usually addressed in terms of relevance judgments. In the respective research tradition relevance is defined as a multi-dimensional concept that cannot be reduced to topicality (i.e., the topical relevance of a piece of information for the current search goal), but is based on a set of different evaluation criteria also reflecting the quality of information (Barry & Schamber, 1998; Bateman, 1998, 1999; Borlund, 2003; Saracevic, 2007a, 2007b; Schamber, 1994). For the most part, however, research in information science dealing with relevance judgments is restricted to information seeking within library catalogues (OPAC) or electronic databases (e.g., Barry & Schamber, 1998; Bateman, 1998, 1999; Borlund, 2003; Hirsh, 1999; Saracevic, 2007a, 2007b; Schamber, 1994; Wang, & Soergel, 1998). Nonetheless, a few studies that addressed the evaluation of Web search results and Web pages yielded the following pattern of results: Besides evaluating the topical relevance (i.e., whether the search result or the Web page matches the search topic), Web searchers also use quality-related evaluation criteria that, for instance, address the *credibility* (e.g., trustworthiness, expertise, author reliability, "officialness") of the information sources or the information therein, the *up-to-dateness* (i.e., currency) of information,

the *intended audience* for whom the information was created, or the *structure and presentation* of information (e.g., clarity, structure, graphics, (un)professional design). Table 1 gives an overview of these studies, which in the following will be described in greater detail. It should be noted, that contrary to most psychological research, these studies provide only descriptive data about users' Web search behavior.

Rieh (2002) examined scholars' (i.e., PhD students and faculty members) relevance judgments during complex Web search tasks (e.g., finding credible information about the disease schistosomiasis, which a fictitious friend was recently diagnosed with). Participants were instructed to search for good, useful, and credible information and to concurrently think aloud, that is, to verbalize everything that came to their mind during their search. Furthermore, in retrospective interviews based on screen recordings of the Web search process participants were asked questions such as "Why did you select this page to look at?" or "Do you believe that this information is good, accurate, current, or correct?". Descriptive results of participants' utterances (i.e., the sum of utterances over all participants) revealed that participants were concerned about information quality and cognitive authority (i.e., credibility) to a substantial extent when they made decisions about which search results to select from the SERPs as well as when they evaluated Web pages. They addressed and evaluated source characteristics, such as source reputation, type of source, and URL domain type, both during search results and Web page evaluation. Furthermore, participants indicated to use their prior domain knowledge to evaluate information quality or credibility. Taken together, more than half of the mentioned criteria for the evaluation of search results were associated with aspects of information quality or credibility, such as accuracy, up-to-dateness, trustworthiness, reliability, "officialness", or authoritativeness. Topicality, however, was the most frequently used (single) criterion. With regard to the evaluation of Web pages, the aspects of information quality or credibility constituted the majority of overall judgments, with topicality having a lower impact. Further less important evaluation criteria were general expectations and aesthetic or affective aspects.

Table 1

Overview of Descriptive Studies reporting on Quality-related Evaluation Criteria

Studies	Criteria found on SERPs	Criteria found on Web pages	Methodology	Participants
Rieh (2002)	• topicality • information quality • credibility	• information quality • credibility • topicality • affective aspects • aesthetic aspects	• evaluation instructions • concurrent thinking aloud • retrospective interviews based on screen recordings	• postgraduates • faculty members
Crystal and Greenberg (2006)	• topicality • research criteria • intended audience • up-to-dateness	• research criteria • topicality • intended audience • up-to-dateness • scientific results or statements • affiliation, authority	• marking • retrospective interviews based on screen recordings	• health information users
Law et al. (2006)	• topicality • quality • up-to-dateness • authority	not assessed	• concurrent thinking aloud • marking	• postgraduates • faculty members
Tombros et al. (2005)	not assessed	• topicality • structure • quality • pictures, physical properties	• concurrent thinking aloud with evaluation instructions	• undergraduates • graduates • academic staff

Crystal and Greenberg (2006) investigated relevance judgments made by health information users who searched for information on a complex medical topic of personal concern (e.g., the risk of car exhaust to cause asthma). Participants were required to examine the first Google SERP returned for their query and to access and read all ten respective Web pages. Their task was to highlight with a mouse every feature they used to evaluate the relevance of the ten search results and the ten corresponding Web pages. Additionally, in a retrospective interview based on screen recordings participants were asked to answer questions regarding their judgments of the usefulness of the search results and the Web pages and the criteria that had influenced their judgments. Data analyses showed that key criteria for evaluating search results were topicality, research criteria (i.e., referring to the research design, methods, or type of study), and up-to-dateness of the content or intended audience of a document. With regard to the evaluation of Web pages, research criteria were the most frequently used criteria, followed by criteria such as topicality, intended audience or up-to-dateness, specific scientific results or statements, affiliation (i.e., reference to an event, an institution, or project), and authority (i.e., name of author, editor, or scientist).

Furthermore, Law, Klobucar, and Pipan (2006) investigated relevance judgments regarding search results provided by a self-developed search engine for learning resources. Unlike Google the search results also included meta-information such as the author, the publisher, number of pages of the documents, and customer reviews. Participants' (postgraduates and faculty members) task was to use the search engine to find learning resources that contributed to their own field of interest. During their Web search participants were required to think aloud and to highlight with a mouse the features that influenced their relevance judgments. Results identified topicality as the evaluation criterion most often applied, being supplemented by some beyond topical criteria like quality, up-to-dateness, and authority.

In another thinking-aloud study, Tombros, Ruthven, and Jose (2005) investigated the criteria Web searchers used to judge the relevance of Web pages (but not of search results) when conducting predefined complex search tasks (e.g., to make an informed decision on the best hi-fi speakers available that fit their budget). The thinking aloud instructions used in this study requested participants (which were undergraduates, postgraduates, and academic staff) to refer to Web page features that helped to

evaluate the documents (e.g. content, layout, source, multimedia content, etc.), and to explain the reasons for which they selected a specific Web page. Tombros et al. (2005) identified the following evaluation criteria used by the searchers to determine the usefulness of Web documents: topicality, structure (e.g., the general layout and clarity of the page), quality (e.g., the authority of the source of information, the up-to-dateness of the information, the overall quality of the Web page), pictures, and physical properties (e.g., accessibility of Web pages, file size, loading time), with first named criteria being used more often. That is, again, study results showed, that topicality was the most frequently mentioned relevance criterion, followed by other quality-related criteria. However, two other descriptive studies, which will be outlined in the following, do not corroborate the findings regarding users' engagement in the evaluation of information quality during Web search (see Table 2 for an overview).

Table 2

Overview of Descriptive Studies mainly reporting on Topicality-related Evaluation Criteria

Studies	Evaluation criteria found on		Methodology	Participants
	SERPs	Web pages		
Savolainen and Kari (2006)	• specificity • topicality • familiarity • insufficient accessibility	• topicality • specificity • insufficient accessibility	• concurrent thinking aloud	• university students • employed peoples
Brand-Gruwel et al. (2008)	• topicality • unspecific criteria	• topicality • unspecific criteria • amount of information	• cued retrospective thinking-aloud (based on eye-movement recordings)	• undergraduate students

Savolainen and Kari (2006) investigated which of 18 topicality- and quality-related evaluation criteria Web searchers used to evaluate and select search results and Web pages. Results indicate that only topical relevance (i.e., topicality and specificity of information), familiarity with a Web site (based on domain or search knowledge), orin case of rejecting search results or Web pages insufficient accessibility of the link

or the content on the page were important evaluation criteria used by the participants (university students and employed people such as engineers or project managers). In this study, participants were instructed to think aloud while searching the Web about self-chosen topics. The authors stressed that other than, for instance, in the study by Rieh (2002) they avoided to ask questions concerning the specific relevance of each search result or Web page, because this "would have significantly decreased the spontaneity of the search process" (Savolainen & Kari, 2006, p. 693).

Similarly, a thinking-aloud study by Brand-Gruwel, Van Meeuwen, and Van Gog (2008) found that undergraduate students searching information about two complex and unknown psychological topics evaluated search results and Web pages mainly based on the connection to task (i.e., the topical relevance) or on the amount of information provided on a Web page. In contrast, evaluation criteria with regard to the quality of information were uttered very rarely. In this study, verbal data were obtained subsequent to the Web search task through the method of cued retrospective reporting based on a replay of eye-movement recordings (cf. Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005); that is, participants were presented a screen recording superimposed with their eye movements and mouse operations that were recorded during Web search and were asked to report retrospectively what they were thinking during task processing. Because eye tracking data reflects visual attention allocation and moment-to-moment cognitive processing during reading- and search-based activities (Rayner, 1998), it is not only suited as a cue for retrospective verbal reporting, but also to directly analyze users' evaluation processes during Web search (see also section 3.4). To sum up, whereas the four information science studies presented in Table 1 reported on searchers' use of quality-related evaluation criteria when accomplishing complex Web search task, the two studies presented in Table 2 revealed mainly topicality-related evaluation criteria to be employed by Web searchers. One possible reason for the inconsistent findings about the use of quality-related evaluation criteria during Web search might be that participants across the studies differed in their prior domain knowledge. In the majority of the studies that found a substantial degree of quality-related evaluation criteria being used, participants seemed to possess a certain amount of domain knowledge regarding the often self-chosen search topics and/or to be experts in information research (e.g., scholars). Another reason for the inconsistent findings might be a methodological one, namely the task instructions or task requirements

given to the study participants (i.e., a contextual variable, cf. Lazonder & Rouet, 2008), as will be outlined in greater detail in the following section.

3.4 The effects of evaluation instructions on the evaluation of information quality during Web search

In most of the reported studies on source evaluations during Web search (or multiple documents reading) participants were instructed beforehand (1) to mention or mark important factors or criteria used to evaluate information (Crystal & Greenberg, 2006; Law et al., 2006; Tombros et al., 2005), (2) to select good or credible information during Web search (Rieh, 2002) or the two most suitable Web pages (Salmerón, Kammerer, et al., 2010), or (3) to rate documents according to given evaluation criteria (Bråten, Strømsø, & Salmerón, 2011; Stanford et al., 2002). Furthermore, in some studies participants were requested to explain their evaluation processes in the form of postsearch interviews including specific evaluation-related questions (Crystal & Greenberg, 2006; MaKinster et al., 2002; Rieh, 2002). The provision of such instructions or task requirements (i.e., a contextual variable, cf. Lazonder & Rouet, 2008), however, may have increased participants' awareness of the necessity of evaluating the information retrieved during Web search, and thus might have activated evaluation processes, in particular regarding the evaluation of information quality, that would not occur spontaneously (cf. Savolainen & Kari, 2006). Indirect evidence for this assumption comes from the studies by Savolainen and Kari (2006) and Brand-Gruwel et al. (2008), which hardly found any quality-related criteria in participants' utterances. These studies did not prompt participants to mention or mark evaluation criteria, to search for high-quality information, or to explain their evaluation strategies. Besides, in their documents model framework, Perfetti et al. (1999) also proposed that task instructions that, for instance, indicate that information may be inconsistent or contradicting across documents, might stimulate participants to direct their attention to source information.

Furthermore, from a methodological perspective, with respect to the research method of collecting thinking-aloud protocols, Ericsson and Simon (1993) claim that only when instructions to think aloud are given in a neutral way by instructing participants to verbalize their thoughts per se without prompting them to explain specific aspects of their behavior, thinking-aloud does not alter participants' course of cognitive

processing. Referring back to Ericsson and Simon's claim, in a methodological study Hertzum, Hansen, and Andersen (2009) examined the effects of neutral thinking-aloud instructions and directed thinking-aloud instructions that included explanation prompts compared to silent task processing on the Web. The tasks were two simple fact-finding tasks and two more complex information seeking tasks on four Web sites representing bookstores and television-channels. To assess participants' information seeking behavior, their eye movements and mouse operations were recorded during search. Furthermore, subsequent to task processing participants' cognitive load was measured. Results showed that as compared to silent task processing the directed thinking-aloud instructions altered participants information seeking behavior more than the neutral thinking-aloud instructions. With the directed thinking-aloud instructions, task completion times were longer, participants showed a more distributed visual exploration of the screen, navigated to more Web pages, and scrolled more frequently within Web pages than in the silent condition. These process measures indicate a more extensive information exploration evoked by the directed thinking-aloud instructions. Furthermore, participants reported a higher cognitive load than in the silent condition. In contrast, with regard to the neutral thinking-aloud instructions, the only two significant differences between the silent condition and the neutral thinking-aloud instructions were an increased cognitive load and a prolonged task processing in the complex tasks when participants were required to think aloud. Thus, although Hertzum et al. (2009) did not directly compare neutral thinking-aloud instructions and directed thinking-aloud instructions, the results of this study indicate that asking participants to explain their actions during Web search stimulated participants more to actively explore information than when just asked to verbalize their thoughts.

However, a study by Britt and Aglinskas (2002) could not find any significant improvements of source evaluation instructions on low prior knowledge students' source evaluations when learning about a historical controversy. In the study half of the college students were given explicit sourcing instructions to attend to information about the authors of the documents and to take into account authors' bias or lack of knowledge while learning about the historical controversy, whereas the other half of the college students received neutral instructions to read the documents to learn about the historical controversy. Results showed that irrespective of the instructions given students' sourcing scores were fairly low. A potential explanation for these

findings is that for students with low prior domain knowledge it would require more than simple instructions to improve their evaluations of source information (Britt & Aglinskas, 2002). Furthermore, as stated by Bråten, Strømsø, and Salmerón (2011) students with low prior domain knowledge might be overwhelmed by the search task itself, which might be even aggravated by certain evaluation instructions.

To conclude, the quality-related evaluation criteria found in the information science studies presented in Table 1 (Chapter 3.3) might have resulted from a combination of explicit evaluation instructions (contextual variable) and a certain level of prior domain knowledge (individual variable). Yet, further research is needed to corroborate this assumption. This was one of the central aims of Study 1 of the present dissertation.

A further concern with regard to the methods used in previous studies investigating searchers' engagement in source evaluations during Web search is the strong focus on consciously accessible verbalized criteria as well as on decisions that lead to overt interactions with the search environment (e.g., mouse clicks). Hence, evaluation processes that go beyond overt actions remain largely undiscovered, and so do quick and unconscious evaluation processes. To unravel these processes that might not show up in overt behavior, eye tracking methodology as a measure of visual attention allocation and moment-to-moment cognitive processing (Rayner, 1998) seems particularly promising to examine Web search and evaluation behavior (e.g., Brand-Gruwel et al., 2008; Granka et al., 2004; Hertzum et al., 2009). For instance, eye tracking allows to reconstruct which search results or Web page characteristics were visually attended to for how long, irrespective from being verbally addressed or not or, in the case of search results, irrespective from being selected or not (cf. Brumby & Howes, 2008; Van Gog, Brand-Gruwel, Van Meeuwen, & Paas, 2008). Moreover, Van Gog, Paas, and Van Merriënboer (2005) argue that especially a combined use of thinking-aloud protocols and eye-tracking data can provide deeper insights into implicit and fine-grained aspects of cognitive processes. Such methodological triangulation (Denzin, 1970, as cited in Scheiter & Van Gog, 2009), that is, the combination of two or more complementary data collection methods, aims at yielding a more complete picture of the processes under investigation. Therefore, in the three studies of the present dissertation both eye-tracking data and thinking-aloud data were recorded to examine participants' evaluation processes during Web search.

3.5   Summary and overview of Study 1

The preceding chapter has illustrated that the current state of the art provides no conclusive evidence whether or under which preconditions users spontaneously engage in evaluations of information quality when searching for science-related issues on the Web. According to information foraging theory (Pirolli, 2007; Pirolli & Card, 1999), judgments of whether to select a search result or not for further processing, are based on proximal cues indicating the topical relevance (i.e., information scent) of the corresponding Web page for a particular search goal. Computational cognitive models based on information scent (e.g., SNIF ACT, CoLiDeS) seem to achieve good model fits without considering the evaluation of information quality as additional criterion. However, it can be claimed that the respective studies only investigated simple fact-finding tasks or tasks for which uncontroversial and consistent information of established quality was provided. The documents model framework (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006), in contrast, addresses more complex tasks about science-related issues for which multiple and potentially contradictory documents of variable quality are provided. The documents model framework predicts that competent readers – as part of the formation of an intertext model – engage in source evaluations by attending to and evaluating specific source characteristics of the documents, such as the document type or the expertise and motives of the document's author, in order to interpret the information provided in the documents in the light of these source characteristics. However, both theoretical considerations (Perfetti et al., 1999; Rouet, 2006) and empirical findings (e.g., Bråten, Strømsø, & Salmerón, 2011; MaKinster et al., 2002; Salmerón, Kammerer, et al., 2010; Stanford et al., 2002) indicate that prior domain knowledge (as an individual variable) plays an important role in making evaluations of information quality during Web search or during studying multiple print documents. Furthermore, evaluation instructions (as a contextual variable) might be a second factor that positively influences the engagement in source evaluations during Web search, thus having resulted in a distortion of spontaneous evaluation processes in the majority of the reported studies on source evaluations during Web search.

Based on these state of affairs, the aim of the first study of this dissertation, which will be presented in the following chapter, was to shed light on laypersons' spontaneous evaluation processes during Web search on a complex medical topic.

More precisely, Study 1 investigated a) what kinds of evaluation processes laypersons engage in on SERPs and on Web pages during Web search on a complex medical issue, and b) what impact evaluation instructions and c) prior domain knowledge have on laypersons' evaluations of information quality. To answer these research questions a combination of different process measures such as eye tracking methodology, log file data (mouse clicks), and verbal protocols were used to examine evaluation processes, thereby aiming at a methodological triangulation.

4 Study 1: The role of evaluation instructions and of prior domain knowledge in evaluating information quality during Web search on medical and health-related issues

4.1 Research questions and hypotheses

Web searching for complex information, such as in the domain of medicine and health care, requires to appropriately evaluating diverse sources of information that vary highly with regard to the quality of information provided. As has been outlined in Chapter 3, information science studies identified different criteria applied by searchers to evaluate Web information (Crystal & Greenberg, 2006; Law et al., 2006; Rieh, 2002; Tombros et al., 2005). According to these studies besides evaluating the topical relevance of information, Web searchers also use other evaluation criteria that address the quality of information, such as the *credibility* (e.g., trustworthiness, expertise, author reliability, "officialness") of the information sources or the information therein, the *up-to-dateness* of information, the *intended audience* for whom the information was created, or the *structure and presentation* of information (e.g., clarity, structure, graphics, (un)professional design).

However, there is one important caveat with regard to these findings, as in these studies prior to the Web search participants were instructed, for instance, to mention or mark important factors provided in the search results and Web pages or criteria used to evaluate information (Crystal & Greenberg, 2006; Law et al., 2006; Tombros et al., 2005), or to select good or credible information (Rieh, 2002). It can be claimed that such explicit evaluation instructions given to the searchers in the reported studies may have increased participants' awareness of the necessity of evaluating the information retrieved during Web search, and thus might have activated evaluation processes, in particular regarding the evaluation of information quality, that would not occur spontaneously (cf. Savolainen & Kari, 2006). Furthermore, participants in these studies seemed to possess a certain amount of domain knowledge regarding the often self-chosen search topics and/or to be experts in information research (e.g.,

scholars). As outlined in Chapter 3 both theoretical considerations (Perfetti et al., 1999; Rouet, 2006) and empirical findings (e.g., Bråten, Strømsø, & Salmerón, 2011; MaKinster et al., 2002; Salmerón, Kammerer, et al., 2010; Stanford et al., 2002) indicate that prior domain knowledge plays an important role in making evaluations of information quality during Web search or during studying multiple print documents.

The first experimental study of this dissertation, thus, tested the general assumption that both explicit evaluation instructions as a contextual variable and prior domain knowledge as an individual variable (according to the conceptual framework proposed by Lazonder & Rouet, 2008) would facilitate Web users' source evaluations on SERPs and on Web pages when engaged in a complex Web search task. It was assumed that explicit evaluation instructions might elicit evaluation processes during Web search that would not occur spontaneously. Accordingly, neutral thinking-aloud instructions (in line with Ericsson & Simon, 1993) were compared to explicit evaluation instructions (i.e., instructions to evaluate search results and Web pages). However, based on the findings by Britt and Aglinskas (2002) that source evaluation instructions did not improve domain novices' source evaluations (cf. Chapter 3.4) it was hypothesized, that explicit evaluation instructions would only improve participants' evaluation processes when a certain amount of prior domain knowledge was given. Students with low prior knowledge on the subject matter of the Web search task instead may be overwhelmed by the search task itself, so that they might have less cognitive resources available to engage in profound source evaluations (cf. Bråten, Strømsø, & Salmerón, 2011). Figure 9 graphically represents this assumed interaction between evaluation instructions and prior domain knowledge on participants' engagement in source evaluations during Web search.
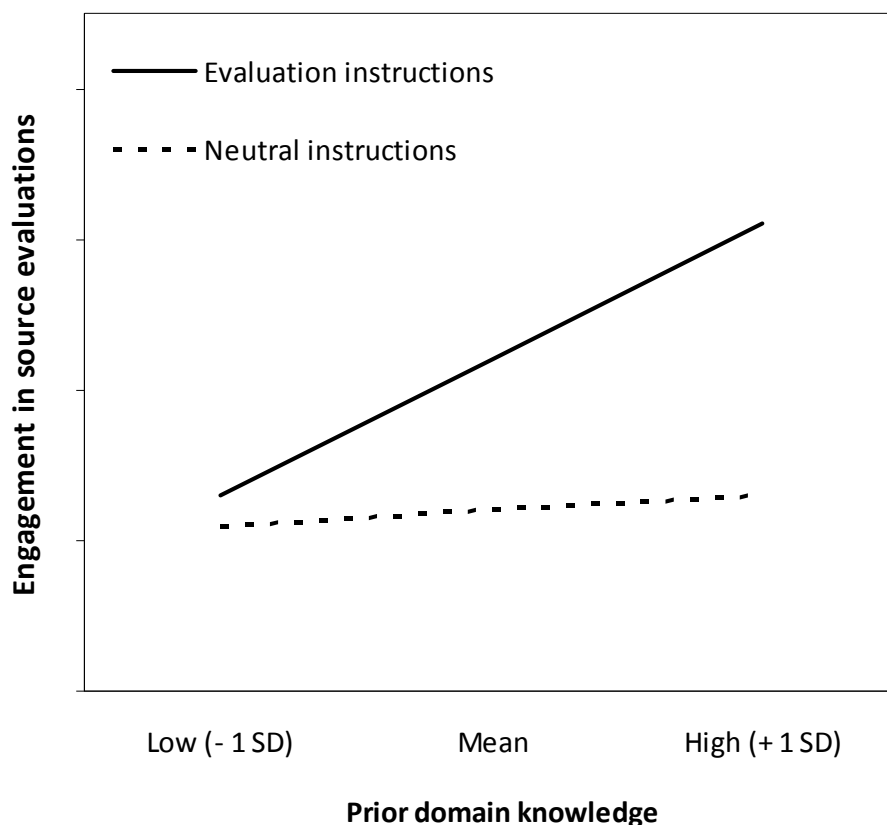
Figure 9. Assumed interaction pattern between evaluation instructions and prior domain knowledge on the engagement in source evaluations.

In order to examine Web users' evaluation processes on SERPs and on Web pages in the present experimental study a combination of different process measures were used. Beyond logging overt interactions with the interface (i.e., the mouse clicks), eye movements were recorded in combination with verbal utterances (thinking-aloud data) to reveal fine grained cognitive processes. While eye tracking methodology, which reflects visual attention allocation and moment-to-moment cognitive processing (cf. Rayner, 1998), seems to be particularly promising to unravel quick and unconscious evaluation processes, verbal protocols provide deeper insights into the "why" of searchers' viewing behavior (Hansen, 1991; Van Gog, Paas, & Van Merriënboer, 2005). Thus, in the present study the focus was on the cognitive processes brought to bear during the first two evaluation phases, that is, during the evaluation of search results and of Web pages (cf. Chapter 2.3). Nevertheless, the third evaluation phase, that is, the evaluation of document collections was not neglected, as in addition to process measures, participants' solution to the

information problem resulting from comparing and integrating information from different information sources was analyzed.

According to these measures, the assumption that the combination of explicit evaluation instructions and a certain amount of prior domain knowledge would facilitate the evaluation of information quality on SERPs and on Web pages resulted in the following more specific hypotheses.

4.1.1 Verbal utterances

It was assumed that explicit evaluation instructions – as compared to neutral instructions – would increase the number of quality-related verbal utterances (e.g., regarding credibility, up-to-dateness, intended audience, or structure and presentation of information) of participants with a certain amount of prior domain knowledge (higher-knowledge participants) both during the evaluation of SERPs and of accessed Web pages. In contrast, for participants with low prior domain knowledge (low-knowledge participants) explicit evaluation instructions were not expected to increase their quality-related verbal utterances, due to participants' lack of cognitive resources to engage in profound source evaluations(Hypothesis 1). For utterances regarding the topical relevance of information (i.e., topicality) no influences of thinking-aloud instructions or prior domain knowledge were expected, because the evaluation of topical relevance (i.e., whether the search result or the Web page addresses the search topic) was considered to be a default process that guides every Web search task.

4.1.2 Eye-tracking data on SERPs

A potential indicator for more thorough evaluation processes on SERPs might be an increased time for which participants attend to the search results. As reported in Chapter 3.1 previous eye-tracking studies – at least for simple fact-finding tasks – have shown that users spontaneously give most attention to the highest-ranked search results on a SERP (e.g., Pan et al., 2007). It was hypothesized that explicit evaluation instructions – as compared to neutral instructions – would increase higher-knowledge participants' attention to lower-ranked search results on a SERP, because the instructions would stimulate them to evaluate all search results by themselves instead of simply relying on the ranking provided by the search engine. In contrast, low-

knowledge participants' visual attention to lower-ranked search results was not expected to be increased by explicit evaluation instructions (Hypothesis 2).

4.1.3 Eye-tracking data on Web pages

More thorough source evaluations on Web pages might be indicated by an increased attention to specific Web page characteristics (i.e., source information) that point to the quality of information, such as author or publisher information, the date of publication, reference information, or the Web site logo. It was hypothesized that explicit evaluation instructions – as compared to neutral instructions – would increase higher-knowledge participants' attention to such kind of source information on Web pages. In contrast, low-knowledge participants' visual attention to source information on Web pages was not expected to be increased by explicit evaluation instructions (Hypothesis 3).

4.1.4   Search result selection

Analogous to the hypothesis of eye-tracking data on SERPs, it was assumed that explicit evaluation instructions – as compared to neutral instructions – would increase higher-knowledge participants' selection of lower-ranked search results from the SERPs, because the instructions would stimulate them to evaluate all search results on their own with regard to information quality instead of selecting links only according to the ranking provided by the search engine. Low-knowledge participants, in contrast, were not expected to select more of the lower-ranked search results when receiving explicit evaluation instructions than when receiving neutral instructions (Hypothesis 4).

4.1.5   Quality of information problem solving

If explicit evaluation instructions stimulated users to evaluate search results and Web pages in the course of their information problem solving, then the quality of the resulting problem solution should improve as well. It was assumed that explicit evaluation instructions – as compared to neutral instructions – would improve higher-knowledge participants' informed decision about the information problem (i.e., a dichotomous decision) as well as their justification of the respective decision. In contrast, low-knowledge participants' solution of the information problem was not expected to be increased by explicit evaluation instructions (Hypothesis 5).

4.2 Method

4.2.1 Participants

Participants were 30 university students (10 male, 20 female; $M = 25.4$ years, $SD = 3.95$, range 21-41 years) from different majors at the University of Tuebingen, Germany; participation was rewarded with either course credit or payment. Because none of the participants were sports or nutritional science students, they were assumed to be laypersons in the content domain of the study (i.e., diets and nutrition). Participants had normal or corrected to normal vision. All participants reported to use Google as their primary search engine.

4.2.2 Experimental design

The study was based on a two-factorial between-subjects design. As a first factor the *thinking-aloud condition*, that is, the type of instructions given to the participants were varied between subjects. Participants received either neutral thinking-aloud instructions (in line with Ericsson & Simon, 1993) without any instructions to evaluate (i.e., neutral instructions, that were assumed to reflect spontaneous evaluation processes) or thinking-aloud instructions including prompts to evaluate (i.e., evaluation instructions).

Neutral instructions to think aloud were worded in line with the standards described by Ericsson and Simon (1993). The instructions were:

> Please think aloud during your Web search, that is, *verbalize everything that comes to your mind.*
>
> Please keep constantly talking from beginning till the end of the task.
> Act as if you were alone, with no one listening, and just keep talking.

In contrast, the evaluation instructions were similar to the instructions used, for instance, by Tombros et al. (2005) or Rieh (2002):

> Please think aloud during your Web search, that is, *mention the evaluation criteria you apply to select search results and to assess Web pages.*
>
> Please keep constantly talking from beginning till the end of the task.
> Act as if you were alone, with no one listening, and just keep talking.

Participants were randomly assigned to the two thinking-aloud instruction conditions, with 15 participants serving in each condition. However, due to technical problems one participant from the group receiving neutral instructions had to be excluded from data analyses.

As a second factor participants' *prior knowledge* on the subject matter of the task (i.e., diets and nutrition) was assessed (see 4.2.3.4 for details) and used as a continuous between-subject factor.

4.2.3 Materials and apparatus

4.2.3.1 Task

A complex and controversial domain characterized by fragile and conflicting evidence was chosen that provided sufficient affordances for searchers to engage in source evaluations (cf. Bråten, Strømsø, & Salmerón, 2009). The task was to achieve an informed decision between low fat and low carb (i.e., carbohydrates) diets with regard to which of the two weight loss methods better promotes a healthy and effective, long-lasting weight loss (cf. the example in Chapters 2 and 3). In line with the method used by Stadtler and Bromme (2007) participants were confronted with a request from a fictitious overweight friend, who wants to lose weight by changing her diet and asks for advice. Participants were asked to conduct a 20-minute Web search regarding this controversial topic in order to decide which of the two weight loss methods they would recommend to their friend. It should be noted that all participants were informed in the task instructions that discrepant information exists about low fat and low carb diets, which should further stress the necessity of source evaluations (cf. Braasch et al., 2010; Rouet et al., 2009, Chapter 3.2.1).

4.2.3.2 Web materials

For their Web search, participants were provided with three prefabricated Google-like SERPs. All three SERPs were accessible by means of a start Web page presenting three hyperlinks with the search terms used to generate these SERPs. The three search terms were "low fat", "low carb", and "low carb + low fat" whereby each of the search terms was used to generate one SERP containing 10 search results (for an example screenshot of the SERP "low fat" see Figure 10). Screenshots of the other two SERPs are provided in Appendix A.

Participants could access 30 Web sites corresponding to the list of search results presented on the SERPs. All search results and Web sites were relevant to the search topic with regard to the content of the information provided. However, the collection of search results and Web sites for each of the three SERPs reflected the given heterogeneity of information sources and their different perspectives and interests with regard to the controversial search topic. All three SERPs included Web sites provided by official institutions (e.g., the German Nutrition Society), journalists (e.g., online magazines), industry and companies (e.g., online shops for nutrition or pharmaceutics), and laypeople (e.g., forum or blog pages). There was an approximately equal distribution of the different types of Web sites between the three SERPs as well as across the positions within a SERP. Accordingly, for each SERP the available information varied largely with regard to contents and quality. However, the sub-collection of information sources that cited scientific evidence mainly favored low fat diets over low carb diets.

The three SERPs and the 30 landing pages of the Web sites were put offline to guarantee a standardized and controlled experimental setting. Hyperlinks within the 30 landing pages could be used whereby in that case further pages of the Web sites were accessed online. It is important to note that none of the participants left any of the 30 Web sites. The layout of the Google-like SERPs was set up close to the original Google layout (as of 2008), but sponsored links (i.e., ads) and the hyperlinks "in cache" and "similar pages" were removed (see Figure 10). Search result links that have already been selected were marked in purple color, whereas not yet selected links were displayed in blue. The Web materials were displayed on a standard 17-inch computer screen and were presented with Microsoft Internet Explorer 6.
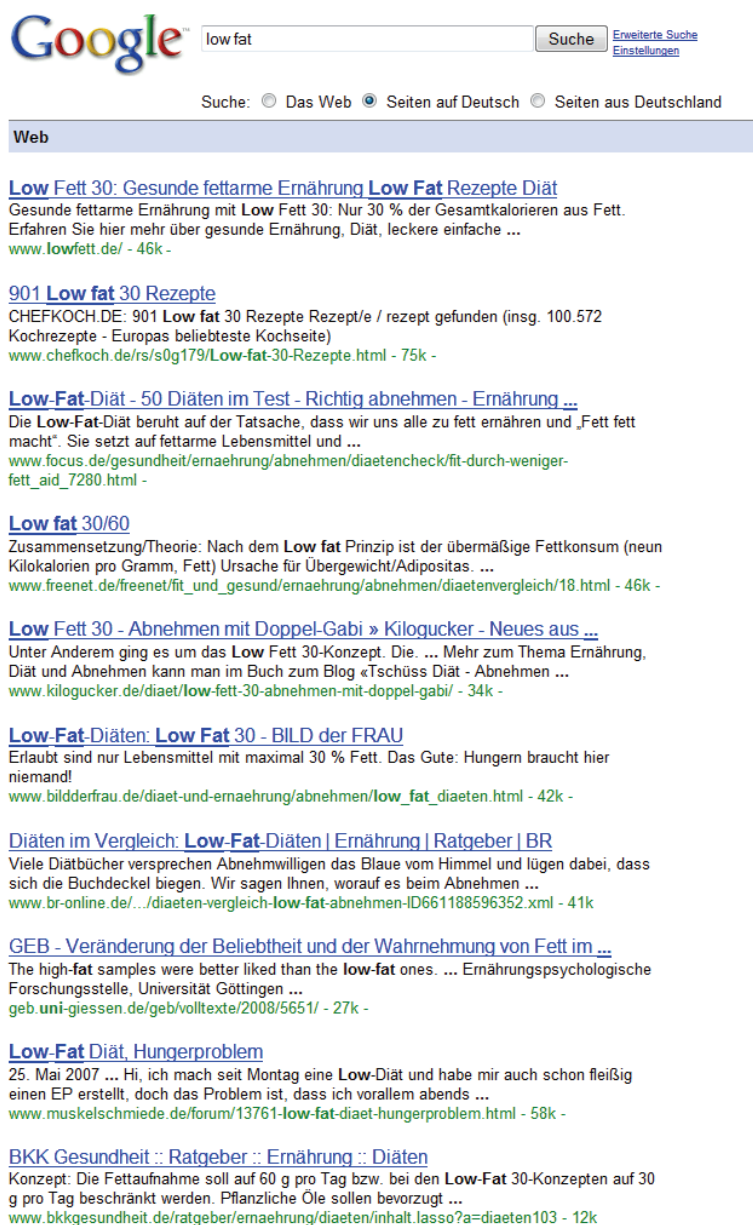
Figure 10. Screenshot of the SERP with "low fat" as keywords.

4.2.3.3 Apparatus

For eye tracking during task processing a 50 Hz Tobii 1750 remote eye tracking system with infrared-cameras built into a 17-inch monitor (www.tobii.com) was used. The Web stimulus recording mode of the ClearView 2.7.1 analysis software was used that captures not only the eye movements, but the entire task performance process (including mouse operations). The minimum fixation duration was set to 100 milliseconds with a fixation radius of 30 pixels (cf. Cutrell & Guan, 2007).

Participants' verbal reports were also recorded by the ClearView software using a standard microphone attached to the PC.

4.2.3.4 Prior domain knowledge measure

To assess participants' prior knowledge on diets and nutrition, in the beginning of the experiment participants were administered a questionnaire with nine statements that had to be rated on 5-point Likert-type response scales ranging from 1 (*totally disagree*) to 5 (*totally agree*); an example item is "I have never heard about the low carb versus low fat controversy". The full list of items as well as means and standard deviations of the items are provided in Appendix B. Cronbach's alpha was .86 for the nine items. Prior domain knowledge scores were normally distributed ($W = 0.97$, $p = .59$) with a mean of 2.44 ($SD = 0.73$) and a range from 1.00 to 4.40. There were no differences between the two experimental conditions regarding participants' prior knowledge on diets and nutrition, $t(27) = -0.03$, $p = .98$ (see Table 3 for means and standard errors).

4.2.3.5 Control variables

Demographics (gender, age) and computer- and Web search experience and skills (3 items, see Appendix C; 5-point scales with 1 = *very low* or *totally disagree* 5 = *very high* or *totally agree*; Cronbach's $\alpha = .74$) were assessed as control variables (see Table 3 for means and standard errors). Analyses of the respective data revealed no differences between the two experimental conditions, that is, for gender, $\chi^2(1, N = 29) = 0.84$, $p = .36$, for age, $t(27) = 0.07$, $p = .95$, and for computer- and Web search experience and skills, $t(27) = -0.35$, $p = .73$.

Table 3

Means (and Standard Deviations) of Prior Domain Knowledge and of Control Variables as a Function of Thinking-aloud Condition

	Neutral instructions ($n = 14$)	Evaluation instructions ($n = 15$)
Age	25.50 (5.01)	25.40 (2.77)
Gender	6 m, 8 f	4 m, 11 f
Prior domain knowledge 1 (low) – 5 (high)	2.44 (0.75)	2.44 (0.73)
Computer- and Web search experience and skills 1 (low) – 5 (high)	3.79 (0.51)	3.87 (0.61)

4.2.4 Procedure

Participants were tested in individual sessions of approximately one hour. Before participants started on the search task, they were administered a computer-based questionnaire to assess prior domain knowledge and control variables. Furthermore, they received some general instructions about the Web search experiment as well as the thinking-aloud instructions according to their experimental condition. Participants were then calibrated on the eye tracking system using a 9-point calibration.

Before working on the main task, participants underwent a training task for approximately five minutes to get acquainted with the thinking-aloud method and the Web search environment. In this training task, they had to conduct a Web search on possible causes and treatments of backache. They were presented three search terms ("backache", "back gym", and "backache + back gym") leading to three Google-like SERPs (with 10 search results each) linked to Web sites. During the training task, participants' thinking aloud was practiced together with the experimenter. In the case that participants did not verbalize their thoughts according to the instructions received the experimenter repeated the instructions and encouraged them to think aloud freely. When the experimenter had the impression that the participants were able to think aloud freely and that they felt comfortable enough with the procedure the training task was finished.

When participants had finished the training task, they were given the instructions for the main task including the request of their fictitious friend to give a recommendation about low carb or low fat diets. Furthermore, participants were again reminded to think aloud during their task performance and to use all three search terms. Eye movements, screen recordings, and concurrent verbalizations were captured during the entire 20 minutes task performance. Whenever participants stopped verbalizing their thoughts, the experimenter reminded them (after approx. 5 seconds) to keep thinking aloud. After 10 minutes the experimenter informed participants that half of the available time was over. Participants were asked to use all three SERPs for their Web search, but were not allowed to generate new SERPs by changing the search terms. Participants could access all Web pages corresponding to the list of search results.

Subsequent to the search task participants were required to decide which of the two weight loss methods they would recommend to their friend and to write a short statement to justify their decision.

4.2.5 Data analyses and dependent variables

4.2.5.1 Coding scheme for thinking-aloud protocols

For the analysis of participants' thinking-aloud protocols a coding scheme was developed that was based on the evaluation criteria found in the information science literature (see Chapter 3.3, Table 1). This scheme was refined and condensed by analyzing data from thinking-aloud protocols of a pilot study. It included the following five evaluation criteria: (a) topicality, (b) credibility, (c) structure, (d) up-to-dateness, and (e) intended audience. The first criterion was content-related, whereas the latter four were quality-related criteria. Short descriptions of these five evaluation criteria are provided in Table 4. Two raters familiar with the search task and the Web materials as well as with the coding scheme scored 30% of the protocols. Interrater reliability computed on this subsample of protocols yielded a Cohen's kappa of .72. Disagreements were resolved through discussion between the raters. One rater scored the remaining protocols. As dependent variables the number of verbal utterances referring to the five different types of evaluation criteria was analyzed for both the search result evaluation and the Web page evaluation. Verbal

utterances expressed across the three SERPs or across the 30 Web sites, respectively, were aggregated in the analyses.

Table 4

Coding Scheme for Thinking-aloud Protocols

Evaluation criteria	Short description
Topicality	The search result or the Web page does (or does not) address the search topic. The search result or the Web page is (or is not) interesting for the search topic because of "what it is about".
Credibility	The information is (or is not) good, valid, credible. The source is (or is not) good, or does (or does not) provide trustworthy, accurate, reliable information, or information provided by experts.
Structure	The structure of the Web page is (or is not) clear, the information is (or is not) well organized.
Up-to-dateness	The information is (or is not) recent or up-to-date.
Intended audience	The material is (or is not) created, for instance, for experts, for laypeople, for patients.

4.2.5.2 Eye-tracking data on SERPs

To analyze participants' eye-tracking data during the evaluation of search results (i.e., search results evaluation phase), on the three SERPs for each of the 10 search results a polygonal "area of interest" (AOI) was defined covering the title, excerpt, and URL of a search result. As no differences were expected between the SERPs eye-tracking data of all three SERPs were aggregated in the statistical analyses. The eye-tracking data recorded during the 20-minute Web search allowed to determine for all AOIs for which amount of time a participant was looking at these areas. Because of the particular interest in participants' evaluation of lower-ranked search results (Hypothesis 2, see section 4.1.2) the *total fixation time on bottom-five search results* (i.e., positions 6 to 10 on each of the three SERPs) was measured as eye-tracking variable during search result evaluation. Total fixation time was defined as the total time for which participants attended to the respective search results on the SERPs. In addition, the *total fixation time on top-five search results* (i.e., positions 1 to 5 on each of the three SERPs) was measured as well. All time data were transformed into seconds for ease of interpretation.

4.2.5.3   Eye-tracking data on Web pages

Furthermore, to analyze participants' eye-tracking data during the evaluation of Web pages, AOIs were defined on all areas on the Web pages that provided source information (i.e., author or publisher information, the date of publication, reference information, or the Web site logo). All AOIs were aggregated for data analyses across the 30 Web sites. As a first eye-tracking variable the *total fixation time on source information* was measured, that is, the total time for which participants attended to source information on the Web pages during their 20-minute Web search. Additionally, the *number of gazes on source information* was measured as a second eye-tracking variable, indicating the frequency with which attention was directed to source information on the Web pages. A single gaze was defined as all successive fixations within the same AOI (cf. Jacob & Karn, 2003).

4.2.5.4 Log file data (mouse clicks)

With regard to participants' decisions which search results to select from the SERPs, analogous to the eye-tracking data the *number of selections (in percent) of the bottom-five search results* (i.e., positions 6 to 10) from the three SERPs was measured as selection variable. In addition, the *number of selections (in percent) of the top-five search results* (i.e., positions 1 to 5) from the three SERPs was analyzed as well (cf. Pan et al., 2007).

4.2.5.5 Quality of decision making and decision justification

Participants' solution to the information problem, that is, the decision to either recommend low carb diets or low fat diets, was analyzed by counting the frequency with which the two diet methods were recommended in each experimental condition. Additionally, participants' statements to justify their decision were rated with respect to their quality on a 3-point rating scale, ranging from 0 (false or no statements), 1 (personal opinions or likes and dislikes without any further argumentation), 2 (fuzzy statements mentioning risks and benefits of one diet method), to 3 points (detailed statement with arguments in favour and against both diet methods). In addition, for exploratory analyses the statements were analyzed with regard to whether or not participants referred to the source of information when justifying their decision.

4.3 Results

An alpha level of .05 was used for the statistical tests reported. Effects were considered marginally significant when p-values were between .05 and .10. Prior domain knowledge was used as a covariate (z-scored) in all analyses. Furthermore, an interaction term between prior domain knowledge and thinking aloud condition was included in the statistical analyses. To analyze the moderating function of prior domain knowledge on effects of the thinking-aloud condition, significant interaction effects between thinking-aloud condition and prior domain knowledge were further examined and graphed using the procedure outlined by Aiken and West (1991). Specifically, simple comparisons of the thinking-aloud instruction conditions (i.e., differences between neutral instructions and evaluation instructions) were computed for different levels of the moderator (i.e., prior domain knowledge), namely for low prior domain knowledge scores (defined as one standard deviation below the sample mean; $M - 1\,SD$) and for high prior domain knowledge scores (defined as one standard deviation above the sample mean; $M + 1\,SD$), using moderated regression analyses. Whereas a median split on prior domain knowledge would result in loss of power due to the truncation of the range of scores into a dichotomous variable, the procedure by Aiken and West (1991) retains the continuous nature of the variable (see also Richter, 2007).

All means and standard errors of the dependent measures reported in the following are corrected for the influence of prior domain knowledge.

4.3.1 Number of verbal utterances

Table 5 shows means and standard errors of the number of verbal utterances related to the five evaluation criteria expressed during the evaluation of search results and of Web pages as a function of thinking-aloud condition (neutral instructions vs. evaluation instructions).

Table 5

Means (and Standard Errors) of the Number of Verbal Utterances related to the Five Evaluation Criteria expressed during the Evaluation of Search Results and of Web Pages as a Function of Thinking-aloud Condition

Evaluation criteria	Evaluation of search results		Evaluation of Web pages	
	Neutral instructions	Evaluation instructions	Neutral instructions	Evaluation instructions
Topicality	3.50 (0.95)	4.60 (0.92)	0.93 (0.58)	1.93 (0.56)
Credibility	1.72 (0.76)	4.52 (0.74)	3.57 (1.01)	9.12 (0.98)
Structure	0.07 (0.07)	0.07 (0.07)	0.93 (0.83)	5.52 (0.80)
Up-to-dateness	0.00 (0.00)	0.00 (0.00)	0.00 (0.16)	0.47 (0.15)
Intended audience	0.07 (0.07)	0.07 (0.07)	0.07 (0.10)	0.27 (0.10)

4.3.1.1 Number of verbal utterances during search result evaluation

With regard to the evaluation of search results, ANCOVAs with thinking-aloud condition (neutral instructions vs. evaluation instructions) as between-subject factor and prior domain knowledge as covariate were calculated for the number of verbal utterances related to the criteria *topicality* and *credibility*. Utterances related to the *structure*, *up-to-dateness*, or *intended audience* were basically not expressed on SERPs (see Table 5). This was not surprising as there were hardly any hints present in the search results about how well-organized or up-to-date the information on a Web page was. Furthermore, verbal evaluations related to the intended audience did not seem to play a role in general for the participants in this study.

For the number of utterances related to *credibility*, the ANCOVA showed significant main effects of thinking-aloud condition ($F(1, 25) = 7.00$, $p = .01$, partial $\eta^2 = .22$) and of prior domain knowledge ($F(1, 25) = 10.82$, $p < .01$, partial $\eta^2 = .30$), as well as a significant interaction between the two factors ($F(1, 25) = 4.12$, $p = .05$, partial $\eta^2 = .14$). In order to probe this interaction, simple comparisons of thinking-aloud condition (neutral instructions vs. evaluation instructions) on different levels of prior domain knowledge were conducted following the procedure outlined by Aiken and West (1991). Results showed that only participants with high prior domain knowledge ($M + 1$ SD) expressed more credibility-related utterances on SERPs when

receiving evaluation instructions than when receiving neutral instructions (ΔM = 4.99, SE = 1.51, $t(25)$ = 3.30, β = .69, $p < .01$). In contrast, for participants with low prior domain knowledge (M - 1 SD), evaluation instructions did not increase participants' number of credibility-related utterances (ΔM = 0.61, SE = 1.51, $t(25)$ = 0.41, β = .09, p = .67). Figure 11 represents this interaction graphically.

Figure 11. Interaction of thinking-aloud condition and prior domain knowledge with regard to the number of verbal utterances on SERPs related to credibility. ** $p \leq .01$

Besides, for the number of *topicality*-related utterances expressed on the SERPs, the ANCOVA showed neither significant main effects of thinking-aloud condition or of prior domain knowledge (both Fs < 1), nor a significant interaction between the two factors ($F(1, 25)$ = 1.38, p = .25).

4.3.1.2 Number of verbal utterances during Web page evaluation

For the evaluation of Web pages verbal utterances related to all five evaluation criteria, namely *topicality, credibility, structure, up-to-dateness,* and *intended audience* were analyzed by means of ANCOVAs with thinking-aloud condition

(neutral instructions vs. evaluation instructions) as between-subject factor and prior domain knowledge as covariate.

As for the evaluation of search results, with regard to the number of *topicality*-related utterances expressed on the Web pages, the ANCOVA showed neither significant main effects of thinking-aloud condition ($F(1, 25) = 1.57$, $p = .22$) or of prior domain knowledge ($F < 1$), nor a significant interaction between the two factors ($F < 1$).

With regard to three of the four quality-related verbal utterances, namely *credibility*, *structure*, and *up-to-dateness*, the ANCOVAs showed significant effects of prior domain knowledge and/or interface. For the number of *credibility*-related utterances expressed on Web pages, there were significant main effects of thinking-aloud condition ($F(1, 25) = 15.47$, $p = .001$, partial $\eta^2 = .38$) and of prior domain knowledge ($F(1, 25) = 4.21$, $p = .05$, partial $\eta^2 = .14$), as well as a significant interaction between the two factors ($F(1, 25) = 8.06$, $p = .01$, partial $\eta^2 = .24$). Simple comparisons revealed that only participants with high prior domain knowledge ($M + 1\,SD$) expressed more credibility-related utterances on Web pages when receiving evaluation instructions than when receiving neutral instructions ($\Delta M = 9.62$, $SE = 2.01$, $t(25) = 4.78$, $\beta = .94$, $p < .001$). In contrast, for participants with low prior domain knowledge ($M - 1\,SD$), evaluation instructions did not increase participants' number of credibility-related utterances ($\Delta M = 1.48$, $SE = 2.01$, $t(25) = 0.73$, $\beta = .14$, $p = .47$). Figure 12 represents this interaction graphically.
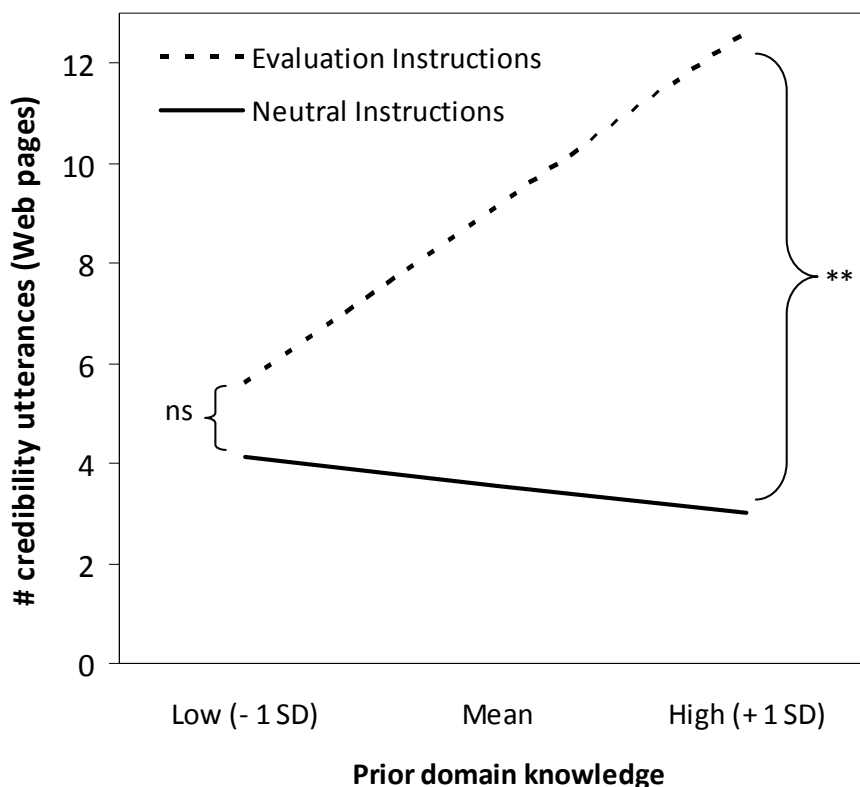
Figure 12. Interaction of thinking-aloud condition and prior domain knowledge with regard to the number of verbal utterances on Web pages related to credibility. ** *p* ≤ .01

Similarly, regarding the number of utterances related to the *structure* of information there was also a significant main effect of thinking-aloud condition ($F(1, 25) = 15.99$, $p < .001$, partial $\eta^2 = .39$), and a marginally significant main effect of prior domain knowledge ($F(1, 25) = 3.19$, $p = .09$, partial $\eta^2 = .11$), together with a significant interaction between the two factors ($F(1, 25) = 5.53$, $p = .01$, partial $\eta^2 = .18$). Simple comparisons revealed that only participants with high prior domain knowledge ($M + 1\ SD$) expressed more utterances related to the structure when receiving evaluation instructions than when receiving neutral instructions ($\Delta M = 7.34$, $SE = 1.64$, $t(25) = 4.48$, $\beta = .91$, $p < .001$). In contrast, for participants with low prior domain knowledge ($M - 1\ SD$) evaluation instructions did not increase participants' number of structure-related utterances ($\Delta M = 1.85$, $SE = 1.64$, $t(25) = 1.13$, $\beta = .23$, $p = .27$). Figure 13 represents this interaction graphically.
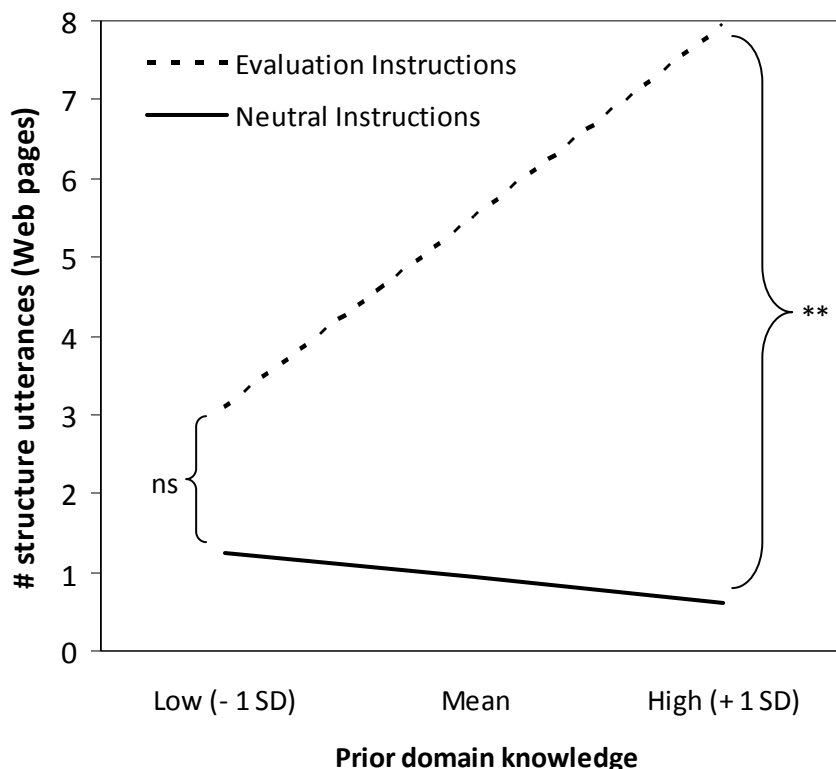
Figure 13. Interaction of thinking-aloud condition and prior domain knowledge with regard to the number of verbal utterances on Web pages related to the structure of information. ** $p \leq .01$

Furthermore, for the number of utterances related to the *up-to-dateness* there was a significant main effect of thinking-aloud condition ($F(1, 25) = 4.55$, $p = .04$, partial $\eta^2 = .15$), with participants receiving evaluation instructions expressing significantly more verbal utterances related to up-to-dateness ($M = 0.47$, $SE = 0.15$) than participants receiving neutral instructions ($M = 0.00$, $SE = 0.16$). However, there was no significant main effect of prior domain knowledge on the number of utterances related to up-to-dateness ($F(1, 25) = 1.63$, $p = .21$) and no significant interaction between the two factors ($F(1, 25) = 1.63$, $p = .21$).

Only for the criterion *intended audience*, there were neither significant main effects of thinking-aloud condition ($F(1, 25) = 1.84$, $p = .19$) or prior domain knowledge ($F < 1$), nor significant interactions between the two factors ($F < 1$) on the number of utterances expressed on the Web pages.

4.3.2 Eye-tracking data

Table 6 shows means and standard errors of the eye-tracking data on SERPs and on Web pages as a function of thinking-aloud condition (neutral instructions vs. evaluation instructions).

Table 6

Means (and Standard Errors) of the Eye-tracking Data on SERPs and on Web Pages as a Function of Thinking-aloud Condition

Eye-tracking data	Neutral instructions	Evaluation instructions
Evaluation of search results (sum of all 30 search results)		
Total fixation time on top-five search results (in seconds)	51.85 (6.37)	54.56 (6.15)
Total fixation time on bottom-five search results	35.90 (6.15)	34.08 (5.94)
Evaluation of Web pages (sum of all 30 Web sites)		
Total fixation time on source information (in seconds)	6.51 (1.51)	8.84 (1.46)
Number of gazes on source information	0.12 (0.69)	2.08 (0.66)

4.3.2.1 Total fixation time on search results

During the 20-minute Web search participants on average visually inspected 92.74% (*SE* = 1.60) of the search results on the SERPs.

With regard to the total fixation time on search results, an ANCOVA with thinking-aloud condition (neutral instructions vs. evaluation instructions) as between-subject factor, search result position (top-five search results vs. bottom-five search results) as within subject-factor, and prior domain knowledge as covariate showed no significant main effect of thinking-aloud condition ($F < 1$), but of prior domain knowledge ($F(1, 25) = 5.96$, $p = .02$, partial $\eta^2 = .19$). The higher participants' prior domain knowledge the longer they fixated on search results, $\beta = .44$, $p = .02$. Furthermore, there was a significant main effect of search result position, $F(1, 25) = 21.09$, $p < .001$, partial $\eta^2 = .46$. Participants fixated significantly longer on top-five search results of the SERPs than on bottom-five search results. However, there were

no significant two-way interactions between thinking-aloud condition, search result position, and prior domain knowledge, and no three-way interaction (all Fs < 1).

4.3.2.2 Total fixation time and number of gazes on source information displayed on Web pages

With regard to the evaluation of Web pages, a MANCOVA with thinking-aloud condition (neutral instructions vs. evaluation instructions) as between-subject factor and prior domain knowledge as covariate was conducted for the total fixation time and number of gazes on source information (i.e., author or publisher information, the date of publication, reference information, and the Web site logo) presented on Web pages. The MANCOVA showed no significant main effect of thinking-aloud condition (F < 1), but a marginally significant main effect of prior domain knowledge (*Pillai's trace* = .20, $F(2, 24)$ = 3.07, p = .07, partial η^2 = .20), and a marginally significant interaction between thinking-aloud condition and prior domain knowledge (*Pillai's trace* = .20, $F(2, 24)$ = 3.04, p = .07, partial η^2 = .20).

Univariate ANCOVAs revealed that these overall effects could be mainly traced back to the number of gazes on source information, with a significant main effect of prior domain knowledge, $F(1, 25)$ = 6.38, p = .02, partial η^2 = .20, and a marginally significant interaction between thinking-aloud condition and prior domain knowledge, $F(1, 25)$ = 3.92, p = .06, partial η^2 = .14. Simple comparisons conducted according the procedure outlined by Aiken and West (1991) revealed that only participants with high prior domain knowledge ($M + 1\ SD$) directed (marginally) more gazes to source information on the Web pages when receiving evaluation instructions than when receiving neutral instructions (ΔM = 9.85, SE = 5.46, $t(25)$ = 1.81, β = .43, p = .08). In contrast, for participants with low prior domain knowledge ($M - 1\ SD$), the type of thinking-aloud instructions had no significant effect on participants' number of gazes on source information (ΔM = -5.55, SE = 5.46, $t(25)$ = -1.01, β = -.24, p = .32). Figure 14 represents this interaction graphically.
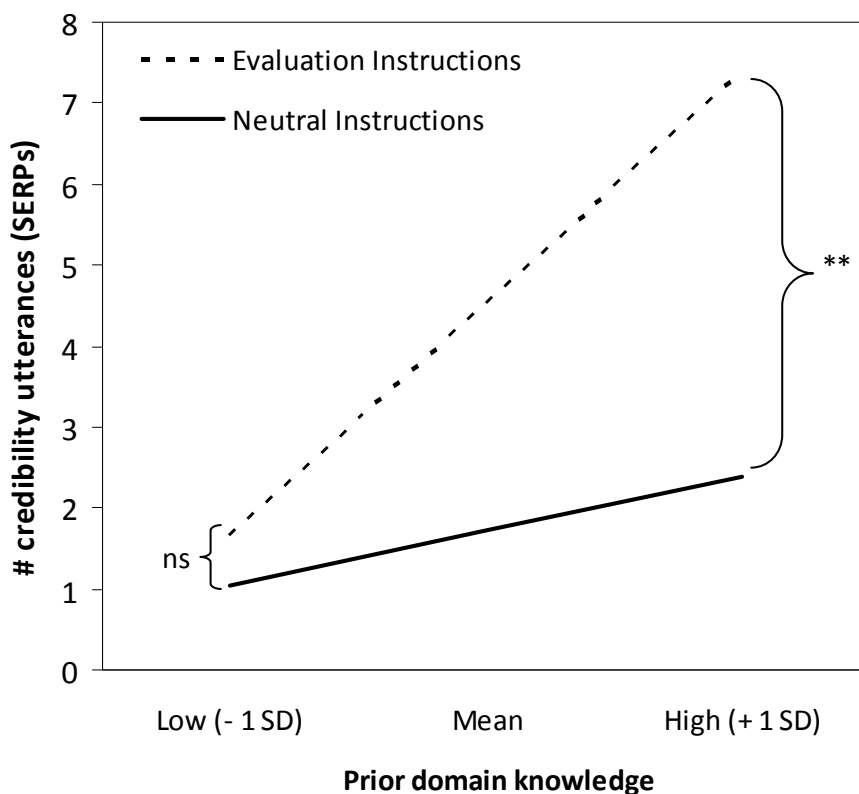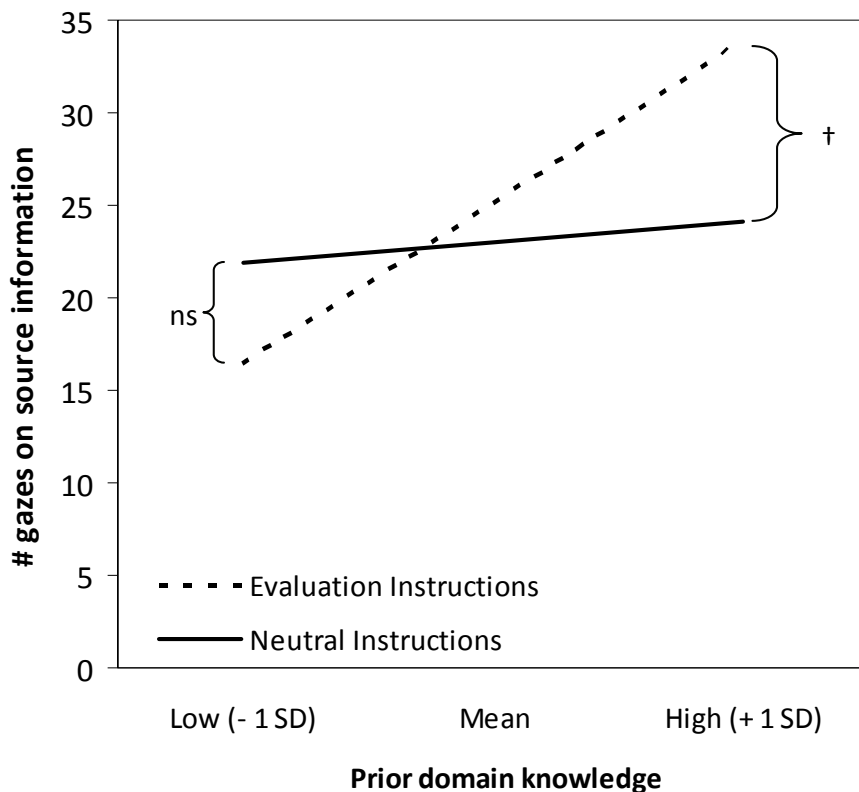
Figure 14. Interaction of thinking-aloud condition and prior domain knowledge with regard to the number of gazes on source information provided on Web pages. † $p \leq$ .10

With regard to the total fixation time on source information, the ANCOVA revealed a marginally significant main effect of prior domain knowledge, $F(1, 25) = 3.08$, $p =$ .09, partial $\eta^2 = .11$, but no significant interaction effect between thinking-aloud condition and prior domain knowledge ($F < 1$). The higher participants' prior domain knowledge, the longer they tended to fixate on source information presented on Web pages, $\beta = .32$, $p = .09$.

4.3.3 Search result selection

Table 7 shows means and standard errors of the percentage of top-five and bottom-five search results selected from the SERPs as a function of thinking-aloud condition.

Table 7

Means (and Standard Errors) of the Selection Data as a Function of Thinking-aloud Condition

Selection data	Neutral instructions	Evaluation instructions
Number of top-five search results selected (in percent)	58.99 (4.46)	55.05 (4.31)
Number of bottom-five search results selected (in percent)	49.03 (5.50)	36.33 (5.32)

During the 20-minute Web search participants on average selected 49.85% ($SE = 3.04$) of the 30 search results from the SERPs.

An ANCOVA with thinking-aloud condition (neutral instructions vs. evaluation instructions) as between-subject factor, search result position (top-five search results vs. bottom-five search results) as within subject-factor, and prior domain knowledge as covariate showed no significant main effects of thinking-aloud condition ($F(1, 25) = 1.87$, $p = .18$) or of prior domain knowledge ($F(1, 25) = 2.81$, $p = .11$) on the number of search results selected. Participants receiving evaluation instructions selected $M = 45.69\%$ ($SE = 4.23$) of the 30 search results and participants receiving neutral instructions selected $M = 54.01\%$ search results ($SE = 4.37$). However, there was a significant main effect of search result position, $F(1, 25) = 17.90$, $p < .001$, partial $\eta^2 = .42$. Participants selected significantly more of the top-five search results of the SERPs ($M = 57.02\%$, $SE = 3.10$) than of the bottom-five search results ($M = 42.68\%$, $SE = 3.82$). Furthermore, there was a significant interaction between prior domain knowledge and search result position, $F(1, 25) = 6.72$, $p = .02$, partial $\eta^2 = .21$. This interaction was explained by a significant positive relationship between prior domain knowledge and the selection of bottom-five search results ($\beta = .38$, $p = .04$), whereas for top-five search results no significant relationship with prior domain knowledge was revealed ($\beta = .04$, $p = .84$). Thus, whereas low-knowledge students selected significantly less bottom-five search results than top-five search results from the SERPs ($\Delta M = -23.28\%$, $SE = 4.77$, $F(1, 25) = 23.19$, $p < .001$, partial $\eta^2 = .48$), high-knowledge students did not differ in their number of top-five and bottom-five search results selected ($\Delta M = -5.38\%$, $SE = 4.77$, $F(1, 25) = 1.25$, $p = .28$). Figure 15 represents this interaction graphically.
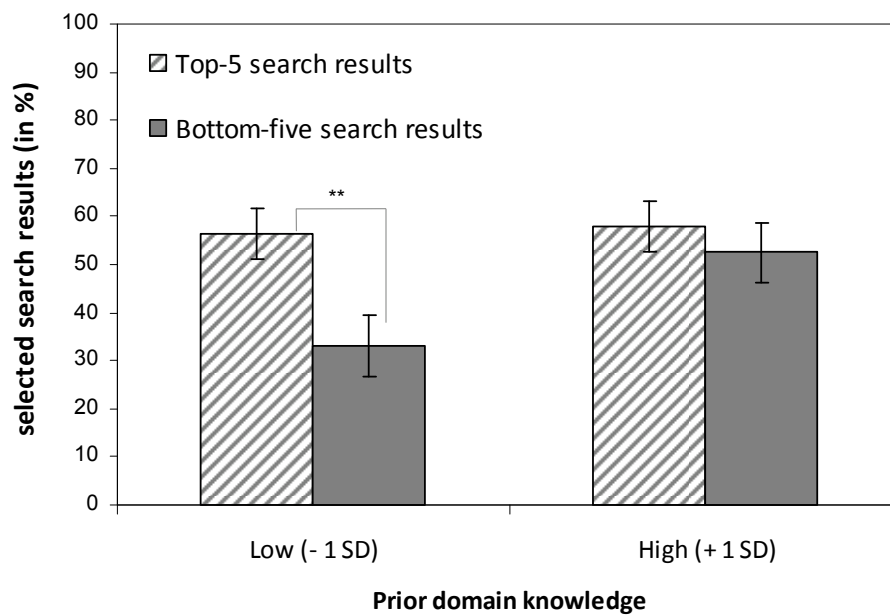
Figure 15. Interaction of prior domain knowledge and search result position (top-five search results vs. bottom-five search results) with regard to the precentage of selected search results. ** *p* ≤ .01.

Moreover, the ANCOVA also showed a significant interaction between thinking-aloud condition and prior domain knowledge, $F(1, 25) = 13.76$, $p = .001$, partial $\eta^2 = .36$. Simple comparisons revealed that whereas participants with high prior domain knowledge ($M + 1\ SD$) tended to select more search results when receiving evaluation instructions than when receiving neutral instructions ($\Delta M = 14.62\%$, $SE = 8.68$, $t(25) = 1.69$, $\beta = .37$, $p = .10$), participants with low prior domain knowledge ($M - 1\ SD$), selected significantly less search results when receiving evaluation instructions than when receiving neutral instructions ($\Delta M = -31.27\%$, $SE = 8.67$, $t(25) = -3.61$, $\beta = -.78$, $p = .001$). Figure 16 represents this interaction graphically.
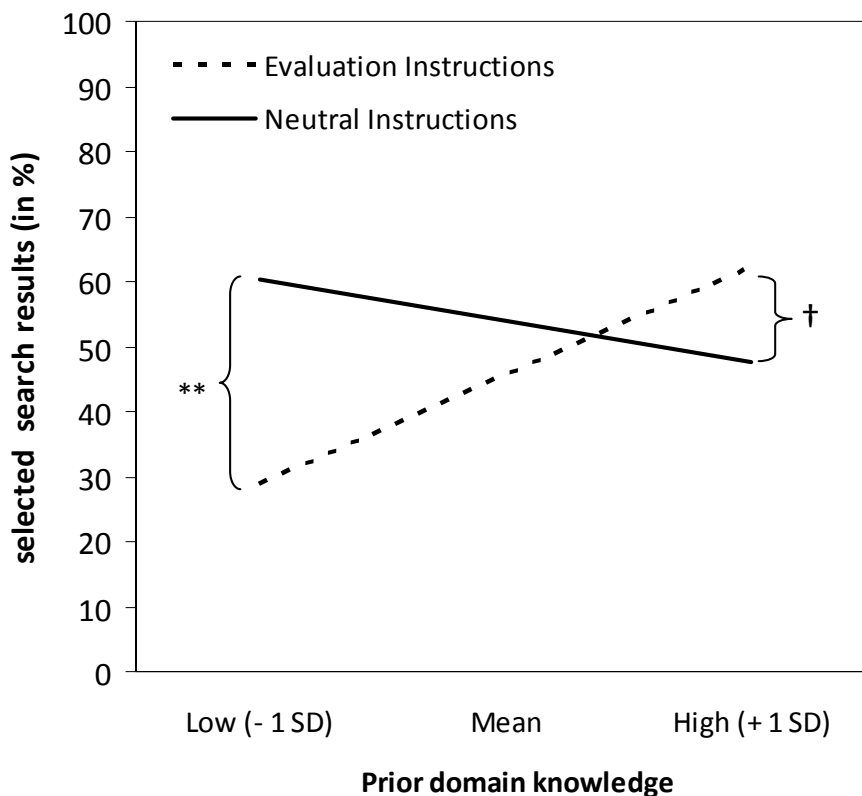
Figure 16. Interaction of thinking-aloud condition and prior domain knowledge with regard to the percentage of selected search results. ** $p \leq .01$, † $p \leq .10$.

Besides, there was neither a significant interaction between thinking-aloud condition and search result position ($F(1, 25) = 1.67$, $p = .21$), nor a three-way interaction between thinking-aloud condition, prior domain knowledge, and search result position ($F < 1$).

4.3.4 Quality of information problem solving

A χ^2-test showed that there was a significant effect of thinking-aloud condition on participants' decision with regard to the information problem, $\chi^2(1, N = 29) = 4.97$, $p = .04$. Whereas in the condition with neutral instructions 10 participants recommended low fat diets and four participants recommended low carb diets, in the condition with evaluation instructions all of the 15 participants recommended low fat diets, which according to the sub-collection of information sources that cited scientific evidence in the present study were the preferable type of weight loss methods. However, this effect was independent of prior domain knowledge. Two of

66

the "low carb recommenders" had a prior knowledge score above average and the other two had a score below average. Furthermore, with regard to the quality of the statements justifying the decision, the results of an ANCOVA with thinking-aloud condition as between subjects factor and prior domain knowledge as covariate showed neither significant main effects of thinking-aloud condition ($F(1, 25) = 1.58$, $p = .22$) or prior domain knowledge ($F(1, 25) = 1.21$, $p = .28$), nor a significant interaction between the two factors ($F < 1$). Participants who had received evaluation instructions achieved $M = 1.73$ points ($SE = 0.25$) for their statements to justify their decision and participants who had received neutral instructions achieved $M = 1.29$ points ($SE = 0.26$).

Finally, exploratory analyses showed no differences between thinking-aloud conditions with regard to whether or not participants referred to the source of information when justifying their decision, $\chi^2(1, N = 29) = 1.01$, $p = .60$. Only three participants (20%) who had received evaluation instructions and one participant (7.14%) who had received neutral instructions referred to the source of information in their statements.

4.4 Discussion

The purpose of the present study was to examine the role of both prior domain knowledge as an individual variable and explicit evaluation instructions as a contextual variable (as classified according to the conceptual framework of Lazonder & Rouet, 2008) in university students' quality-related evaluation processes during Web search for a complex health-related issue. Based on theoretical considerations and former research on evaluation processes, it was assumed that, given that the student possessed a certain amount of prior domain knowledge, explicit evaluation instructions (i.e., to mention during Web search the evaluation criteria one applies to select search results and to assess Web pages) would increase the evaluation of information quality during Web search. Therefore, the present study compared standard (i.e., neutral) thinking-aloud instructions to explicit evaluation instructions. Furthermore, participants' prior knowledge on the search topic was assessed.

As expected, interaction effects of thinking-aloud condition and prior domain knowledge on the evaluation of information quality were found for verbal utterances (Hypothesis 1) as well as for eye-tracking data on Web pages (Hypothesis 3) and for

search result selection (Hypothesis 4), but not for eye-tracking data on SERPs (Hypothesis 2) or for participants' information problem solving (Hypothesis 5). For eye-tracking data on SERPs, however, a main effect of prior domain knowledge was found, and for participants' information problem solving a main effect of thinking-aloud condition. The following two sections provide a detailed discussion of the results of this study regarding the evaluation of search results and the evaluation of Web pages.

4.4.1 Influences of evaluation instructions and prior domain knowledge on the evaluation of search results

In line with Hypothesis 1, during search-result evaluation the interaction effect of thinking-aloud condition and prior domain knowledge was shown for the number of credibility-related verbal utterances. Students with higher prior domain knowledge who received explicit evaluation instructions (i.e., to mention the evaluation criteria they apply to select search results and to assess Web pages) verbally reflected more on the credibility of information sources when they inspected the SERPs than students with higher prior domain knowledge who received neutral thinking-aloud instructions (i.e., to verbalize everything that comes to their mind). In contrast, as expected, explicit evaluation instructions did not increase low-knowledge students' number of credibility-related verbal utterances during search result evaluation. Thus, irrespective of the instructions given, students with no or little prior domain knowledge only rarely referred to the credibility of information sources when evaluating search results. Verbal utterances regarding structure, up-to-dateness, or intended audience were basically not expressed on SERPs. The most frequently expressed utterances on SERPs were related to the topical relevance of the search results. As expected these utterances were unaffected by prior domain knowledge or by the instructions given.

However, contrary to Hypothesis 2, explicit evaluation instructions did not increase high-knowledge students' attention to lower-ranked search results on the SERPs. Irrespective of thinking-aloud condition or prior domain knowledge participants directed more attention to the top-five search results on the SERPs than to the bottom-five search results. Nonetheless, the higher students' prior knowledge on diets and nutrition the longer they fixated on all search results on the SERPs, indicating a more thorough evaluation of the results provided by the search engine.

With regard to the selection of search results from the SERPs in order to access Web pages, the higher students' prior domain knowledge the more search results with lower rankings they selected from the SERPs. Thus, whereas low-knowledge students selected more top-five search results than bottom-five search results from the SERPs, high-knowledge students did not differ in their number of top-five and bottom-five search results selected during Web search. Contrary to Hypothesis 4, this effect of prior domain knowledge, however, was independent of the instructions given. Furthermore, students with higher prior domain knowledge tended to select more search results from the SERPs when receiving explicit evaluation instructions than when receiving neutral thinking-aloud instructions. In contrast, for students with low prior domain knowledge the reversed pattern was true. A potential reason for the finding that students with low prior domain knowledge selected less search results when explicit evaluation instructions were given than with neutral thinking-aloud instructions, might be that they were overwhelmed by the search task and the evaluation instructions (cf. Bråten, Strømsø, & Salmerón, 2011).

4.4.2 Influences of evaluation instructions and prior domain knowledge on the evaluation of Web pages

As for the evaluation of search results, for the evaluation of Web pages the hypothesis that explicit evaluation instructions would increase the number of quality-related verbal utterances of searchers with a certain amount of prior domain knowledge was confirmed (Hypothesis 1). Students with higher prior domain knowledge reflected more on the credibility or structure of a Web page or the information therein when receiving explicit evaluation instructions than when receiving neutral thinking-aloud instructions. In contrast, as expected, explicit evaluation instructions did not increase the number of respective verbal utterances of low-knowledge students. Merely regarding the up-to-dateness of information on Web pages, explicit evaluation instructions increased students' utterances irrespective of their prior domain knowledge. Though, verbal reflections on up-to-dateness as well as on intended audience were extremely rare during Web page evaluation. Verbal utterances related to the topical relevance of information were also rather seldom and were not affected by prior domain knowledge or by the instructions given.

Furthermore, in line with Hypothesis 3, the interaction effect of thinking-aloud condition and prior domain knowledge was shown for the number of gazes directed

to specific Web page characteristics that point to the quality of information, such as author or publisher information, the date of the publication, reference information, or the Web site logo. During their Web search, students with higher prior domain knowledge directed more gazes to such kind of source information displayed on the Web pages when receiving explicit evaluation instructions than when receiving neutral thinking-aloud instructions. In contrast, low-knowledge students' attention to source information on Web pages was not affected by the type of instructions given. With regard to the time spent fixating on the respective Web page characteristics, effects of prior domain knowledge were independent of the thinking-aloud condition. The higher participants prior domain knowledge the longer they tended to fixate on source information displayed on the Web pages. Thus, eye-tracking results indicate that explicit evaluation instructions increased the frequency with which high-knowledge students directed their attention towards source information displayed on Web pages, even though the amount of time spent visually inspecting the source information was not increased.

Finally, explicit evaluation instructions positively influenced participants' solution of the information problem. In this study, all participants with explicit evaluation instructions decided to recommend low fat diets, whereas when neutral thinking-aloud instructions were given 10 participants recommended low fat diets and 4 participants recommended low carb diets. As the sub-collection of sources in the present study that cited scientific evidence mainly favored low fat diets over low carb diets the decision of the participants who received explicit evaluation instructions can be seen as superior. However, contrary to Hypothesis 5 this effect was independent of prior domain knowledge. Furthermore, explicit evaluations instructions in the present study did neither enable students – not even those with higher prior domain knowledge – to better justify their decision in favor or against low carb or low fat diets, nor to refer more to the source of information when justifying their decision. A potential reason for the lacking effects of both prior domain knowledge and explicit evaluation instructions on the quality of students' statements to justify their decision on weight loss methods might be that writing a one- or two-sentence statement to justify one's decision was too narrow a task to find differences in source evaluations. Therefore, in Study 2 and Study 3 of this dissertation as product of the search task participants were required to list pro and con arguments (Study 2) or to write an argumentative essay (Study 3) about a complex medical issue. Besides, it has to be

acknowledged, that students' prior domain knowledge in general was quite low in the present study. It might be assumed that when investigating participants with a broader range of levels of prior domain knowledge larger differences in evaluation processes and search outcomes could be found.

4.4.3 Conclusions

In summary, the results of the present study showed that, during both the evaluation of search results and of Web pages, explicit evaluation instructions not only increased students' quality-related verbal utterances, but also changed their processing strategies as indicated by log files and eye-tracking data, given that students possessed a certain amount of prior knowledge on the search topic. Furthermore, explicit evaluation instructions at least to some extent improved students' information problem solving. Thus, findings from information science studies (Crystal & Greenberg, 2006; Law et al., 2006; Rieh, 2002; Tombros et al., 2005) in which such evaluation instructions were used, can be assumed to not reflect the evaluation criteria laypersons spontaneously apply during Web search. Moreover, results of the present study indicate that, irrespective of the instructions given, students with no or little prior knowledge on the subject matter of the search only rarely evaluated the quality of information and to a great extent simply relied on the ranking provided by the search engine.

Yet, from a methodological point of view it might be criticized that even the standard thinking-aloud instructions that in the present study were used as a control condition for the explicit evaluation instructions are still not very close to a natural search situation. The procedure of the experiment in both conditions required participants to verbalize their thoughts concurrently to their search process, which might have interfered with participants' cognitive processes. The use of obtrusive measurements like concurrent thinking aloud may have reactively influenced participants' search process itself in a positive or negative way - depending on participants' cognitive prerequisites. For instance, it is conceivable that even neutral thinking-aloud instructions might have stimulated participants with higher prior knowledge to process the materials more elaborately, or that the same instructions might have additionally increased low-knowledge participants' cognitive burden during task processing (cf. Hertzum et al., 2009; Van Gog et al., 2008). To investigate Web users' natural search behavior more unobtrusively, in Study 2 and 3 of this

dissertation, therefore, eye tracking will be combined with cued retrospective verbal utterances by presenting participants with a screen recording of their task processing superimposed with their eye movements and mouse operations subsequent to the search task and asking them to report retrospectively what they were thinking during information seeking (cf. De Koning, Tabbers, Rikers, & Paas, 2010; Jarodzka, Scheiter, Gerjets, & Van Gog, 2010; Van Gog, Paas, Van Merriënboer, & Witte, 2005).

From an educational point of view the findings of the present study indicate that simple evaluation instructions as used in the present study have the potential to stimulate searchers to engage in source evaluations during Web search, given that they possess a certain amount of prior knowledge on the subject matter of the search task at hand. However, the present findings also corroborate those of Britt and Aglinskas (2002) that for searchers with low prior domain knowledge simple prompts to apply evaluation criteria or to attend to source information on SERPs and on Web pages are not sufficient to improve their evaluations of information quality during Web search. Yet, as individuals who search for medical and health information on the Web, for example, to learn about potential risks and benefits of different medical treatments, often possess only low prior knowledge on the respective medical topic (cf. Stadtler & Bromme, 2007, 2008) adequate instructional measures that enhance low-knowledge searchers' source evaluation strategies during Web search are of particular relevance. In recent years, a number of instructional measures have been developed and evaluated that seem promising to stimulate and/or train (low-knowledge) searchers to critically and adequately evaluate the quality of information sources and the information they contain.

For example, Stadtler and Bromme (2007, 2008) developed a browser-based support tool called *met.a.ware* that provides users with a structured note-taking facility to systematically store information found during Web search. Moreover, the tool provides specific prompts to evaluate the information retrieved from the Web according to its credibility. Every time users paste information from a Web page into the note-taking facility, they are required to indicate the source of information, as well as to rate the author's expertise, the bias of information, and their confidence in the information on 5-point scales. In an experimental study, in which university students with low prior domain knowledge had to conduct a Web search about a complex medical topic (i.e., cholesterol), Stadtler and Bromme (2007, 2008) showed

that evaluation prompts provided by the *met.a.ware* tool improved students' knowledge about sources as well as increased their number of arguments relating to the source of information when justifying credibility judgments, as compared to a control group not receiving these prompts. However, it should be noted that *met.a.ware* does not address the evaluation of search results.

In addition to such kind of indirect support measures, other instructional measures include direct interventions to teach students how to engage in source evaluations. For example, Britt and Aglinskas (2002) developed the Sourcer's Apprentice, a computer-based application that trains students to identify and evaluate source information (e.g., about the author's expertise and possible motives, the document type, or the publication date) while reading multiple documents about a historical controversy. Two central components of the Sourcer's Apprentice are (1) an interactive tutorial that directly and explicitly informs students about various source features, and (2) a practice environment where students practice their sourcing skills by filling in structured note cards for each document with slots for the author's name, position, or motives, etc. A series of evaluation studies indicate that the Sourcer's Apprentice as compared to regular classroom activities or textbook reading improved high-school students' sourcing skills in the context of researching an unknown historical controversy.

Wiley et al. (2009) developed the SEEK training, a one-hour instructional unit for undergraduate students that focuses on how to evaluate the quality of Web-based information sources and the information they contain. The instructional unit consisted of three parts: first, students were provided with a description about how to evaluate information quality (e.g., by determining the author's expertise and motives, by comparing information across information sources, etc.). Second, students were asked to apply these criteria to the evaluation of six Web sites on a health-related topic. Third, students had to rank-order the six Web sites according to their credibility. After they had assigned their rankings, they were provided with feedback about how 10 experts had ranked the same Web sites. In an experimental study, Wiley et al. (2009) requested undergraduate students to conduct a Web search about possible causes of volcanic eruptions. Results showed that, as compared to a control group, students who had received the SEEK training were more capable to differentiate between reliable and unreliable information sources (as measured by rank-ordering search results on a SERP after having explored all corresponding Web

sites) and to articulate their reasons for their reliability evaluations. Furthermore, students who had received the SEEK training wrote sounder essays about the search topic than students without the SEEK training.

In contrast, more comprehensive Web search training programs for younger students (ages 12-15), which consist of multiple lessons to foster students' Web search and evaluation skills, such as the CIS-WEB (Competent Information Search in the World Wide WEB) training by Gerjets and Hellenthal-Schorr (2008) or two training programs by Walraven, Brand-Gruwel, and Boshuizen (2010) have met with only moderate success. For example, results of an evaluation study by Walraven et al. (2010) indicated that both of their training programs – the "high road" program that trained evaluation skills according to the different phases of the information-problem solving process and the "rich representation" program that conveyed relevant evaluation criteria for the evaluation of search results and Web pages – improved students' abilities to indicate Web page features that point to the quality of information. However, the training programs did not improve students' abilities to distinguish between reliable and unreliable search results. Thinking-aloud data obtained for a subgroup of participants furthermore revealed that after the trainings students still mainly focused on evaluating the topical relevance of the search results on the basis of the title and the page excerpt of the search results.

To sum up, the abovementioned instructional measures seem to have positive effects on students' source evaluation processes on Web pages. Yet, with regard to the evaluation of search results except for the SEEK training by Wiley et al. (2009) the instructional measures either did not address the evaluation of search results (e.g., Stadtler & Bromme, 2007, 2008) or achieved no respective improvements (e.g., Walraven et al., 2010). Furthermore, trainings or the use of computer tools are time-consuming and cannot be provided to everyone.

Therefore, from a theoretical point of view, it is an interesting question whether there exist other, more domain-independent preconditions under which even searchers with low prior domain knowledge engage in source evaluations during Web search on a controversial medical issue, without providing them with trainings or prompting tools like *met.a.ware*. Two promising candidates for the preconditions might be searchers' personal epistemology as an individual variable and the search interface as a resource variable (according to the terminology by Lazonder & Rouet, 2008).

Accordingly, the following chapter theoretically analyzes the role of these two variables in making evaluations of information quality. Subsequently, Chapters 6 and 7 describe two experiments (Study 2 and Study 3) on the effects of personal epistemology and of the search interface on low-knowledge searchers' source evaluations during Web search on a complex medical issue.

5   The role of personal epistemology and of the search interface in evaluating information quality during Web search

Results of Study 1 indicate that university students, who were laypersons in the field of medicine and health care, spontaneously engaged in source evaluations only to a rather limited extent when conducting a Web search on a controversial health-related issue. In particular, students with no or little prior knowledge on the subject matter of the search task only rarely evaluated the quality of information and to a great extent simply relied on the ranking provided by the search engine. Thus, instead of trying to "separate the wheat from the chaff" on their own, they seemed to count on the search engine to do "this job" for them. However, ranking algorithms of popular search engines, such as Google, are mainly based on topical relevance and popularity of the Web pages (Cho & Roy, 2004), but not on information quality. Thus, the results lists returned by search engines usually present different types of Web sources such as articles from scholarly journals or newspapers, or official reports, forum discussions or blog postings, and commercial information, interspersed with each other. Moreover, in many cases popular commercial or social Web sites (e.g., shops or forums) that may be doubtful with regard to their authors' motives or expertise fit exactly to search terms entered by users, so that they are listed among the highest-ranked search results on the first SERP. Thus, even the information contained in the top search results of a SERP might turn out to be subjective, unscientific, one-sided, or commercially biased (Lewandowski, 2011) rather than scientifically sound and objective. Accordingly, from a normative perspective Web searchers should critically evaluate the search results in terms of information quality (cf. Taraborelli, 2008) – especially when dealing with complex, controversially discussed issues such as the effectiveness of specific weight loss methods or medical treatments.

Therefore, the remainder of this dissertation aimed at identifying factors that potentially facilitate or hamper spontaneous source evaluations on SERPs when searching on the Web for complex medical topics, for which prior knowledge is low. A central goal was to examine whether there exist preconditions under which even searchers with low prior domain knowledge engage in source evaluations during

Web search on a controversial medical issue, without providing them with trainings or educational computer tools.

As outlined in the end of the previous chapter, educational computer tools or trainings such as *met.a.ware* (Stadtler & Bromme, 2007, 2008), the Sourcer's Apprentice (Britt & Aglinskas, 2002), or the SEEK training (Wiley et al., 2009) are designed to stimulate searchers to critically evaluate the quality of information sources and the information they contain. Whether searchers – without such prompting tools or trainings – are aware of the necessity to critically evaluate the quality of information, however, might also depend on their *personal epistemology* or *epistemic beliefs*, that is, their conceptions about the nature of knowledge and knowing (Hofer 2004; Hofer & Pintrich, 1997). Searchers' epistemic awareness of the varying information quality on the Web and the associated epistemic challenges involved in managing the wealth and heterogeneity of information sources, for example, might be an important precondition for source evaluations; in particular for searchers with no or little prior knowledge on the search topic who are not able to determine by the content whether information is accurate or not.

However, holding a particular belief, for example that the quality of Web information varies greatly, does not automatically imply that a searcher will necessarily engage in critical source evaluations during Web search (cf. Bråten, Britt, et al., 2011). Rather, whether an epistemic belief is activated during Web search might also depend, for example, on the search interface (e.g., the interface presenting search results retrieved by a search engine) a Web user has at his or her disposal to evaluate and select potentially relevant Web pages. A search interface that presents search results in a rank-ordered list format and that displays only sparse and non-salient source information for the search results returned by the search engine might not sufficiently activate epistemic doubts with regard to the quality of Web information. For this type of interface, thus, it seems plausible that in particular searchers with low prior domain knowledge who might have less cognitive resources available to engage in profound source evaluations (cf. Bråten, Strømsø, & Salmerón, 2011), tend to base their selection decisions about which Web pages to access for further inspection on the ranking position of the search results (cf. MaKinster et al., 2002), instead of taking over the responsibility for evaluating and selecting search results on their own in terms of information quality.

The following sections outline theoretical considerations as well as empirical findings on the role of epistemic beliefs as an individual variable (Chapter 5.1) and on the role of the search interface as a resource variable (Chapter 5.2) in making evaluations of information quality during Web search or during studying multiple print documents.

5.1 The role of epistemic beliefs in evaluating information quality

In the last two decades, in the field of educational psychology people's beliefs about the nature of knowledge and knowing, that is, their personal epistemology or epistemic beliefs, have received major attention with regard to their role in learning and comprehension (Hofer, 2004). Research indicates (for an overview see Opfermann, 2008) that epistemic beliefs are of particular importance a) when working on complex learning tasks about controversially discussed, ill-structured problems (e.g., Schraw, Dunkle, & Bendixen, 1995) and b) when working with hypermedia technologies such as the WWW, where learning and information seeking affords a higher degree of self-regulation and evaluation strategies than in traditional learning environments (Bendixen & Hartley, 2003; Jacobson & Spiro, 1995). Thus, it seems plausible that epistemic beliefs also play an important role in source evaluations during Web search on complex science-related topics. Before reviewing first empirical evidence on this relationship in section 5.1.3, section 5.1.1 provides a conceptual framework of personal epistemology. Furthermore, on the basis of a theoretical model proposed by Bråten, Britt, et al. (2011) section 5.1.2 describes how epistemic beliefs are related to readers' engagement in source evaluations when reading multiple documents.

5.1.1 A multidimensional and multilayered framework of personal epistemology

According to Bråten, Britt, et al. (2011) the currently most widely accepted conceptualization of epistemic beliefs, at least in educational psychology, is the multidimensional framework of personal epistemology by Hofer and Pintrich (1997). Following Hofer and Pintrich, epistemic beliefs can be defined on four continuous dimensions that range from naïve to sophisticated beliefs. The first two dimensions

refer to the nature of knowledge (i.e., what one believes knowledge is) and concern the *simplicity of knowledge* (i.e., the degree to which knowledge is considered as simple or complex) and the *certainty of knowledge* (i.e., the degree to which knowledge is considered as absolute or tentative and evolving). The remaining two dimensions refer to the nature of knowing (i.e., how one comes to know) and concern the *source of knowledge* (i.e., the degree to which knowledge is considered as transmitted by external authorities or as constructed by the self), and the *justification for knowing* (i.e., the beliefs how knowledge claims can be warranted). Thus, in contrast to an earlier conceptualization of epistemic beliefs by Schommer (1990) that not only comprised beliefs about knowledge and knowing (i.e., certain knowledge, simple knowledge, and omniscient authority), but also about learning and intelligence (i.e., quick learning and fixed ability), the conceptualization by Hofer and Pintrich (1997) only represents the core structure of personal epistemology. However, the terminology of naïve or low-level beliefs and sophisticated or high-level beliefs used by Hofer and Pintrich (1997) and Schommer (1990) to describe the endpoints of the dimensions recently has been criticized, because a particular belief might be adaptive in one context, but not in another one (e.g., Hofer & Sinatra, 2010; Strømsø & Bråten, 2010). Therefore, in line with a suggestion by Bråten, Britt, et al. (2011) in the remainder of the present dissertation epistemic beliefs were described in more neutral terms (e.g., as beliefs in certain knowledge or in tentative and evolving knowledge).

Besides discussions about the conceptualization of epistemic beliefs, another controversially discussed issue concerns the domain-generality or domain-specificity of epistemic beliefs (for reviews see Buehl & Alexander, 2001; Muis, Bendixen, & Haerle, 2006). Buehl and Alexander (2001) as well as Muis et al. (2006), however, concluded that it is not a question of "either-or", but that epistemic beliefs are multilayered, that is, that students hold both beliefs about knowledge and knowing in general (domain-general beliefs) and beliefs about knowledge and knowing in specific domains such as medicine or physics (domain-specific beliefs) or with regard to specific topics such as climate change (topic-specific beliefs). As domain-specific or topic-specific beliefs have been shown to be related to the level of domain or topic knowledge (for a review see Bromme, Kienhues, & Stahl, 2008), for Study 2 and 3 of the present dissertation that examined searchers with low prior domain knowledge, it was decided to assess epistemic beliefs that are independent of the search domain.

In addition to considerations about the content-related specificity of personal epistemology, epistemic beliefs can also be considered specific with regard to a particular information technology, such as the Web. For example, Bråten and Strømsø (2010) claim that when examining learning and information search on the Web "[...] epistemic belief measures should be tailored to this information technology because it allows for new ways of presenting knowledge and new ways of knowing" (p. 95). Therefore, they developed the Internet-Specific Epistemological Questionnaire (ISEQ), that measures *Internet-specific* epistemic beliefs, that is, beliefs about the nature and use of information found on the Web (cf. Strømsø & Bråten, 2010; Bråten, Strømsø, & Samuelstuen, 2005). Due to its specificity for Web search activities, this questionnaire was used in Study 2 and 3 of this dissertation.

Previous research provides amounting evidence that epistemic beliefs affect the processing and comprehension of multiple print or hypertext documents that contain conflicting information and opposing perspectives on a certain issue (e.g., Bråten & Strømsø, 2010b; Bråten, Strømsø, & Samuelstuen, 2008; Jacobson & Spiro, 1995; Pieschl et al., 2008; Strømsø, Bråten, & Samuelstuen, 2008). This led Bråten, Britt, et al. (2011) to propose a model that integrates epistemic beliefs into the documents model framework (originally proposed by Perfetti et al., 1999; see Chapter 3.2), specifying how and why epistemic beliefs might facilitate or impair the construction of a documents model. The following section will outline the assumptions of this theoretical model about the relationship between epistemic beliefs and readers' engagement in source evaluations as part of the formation of an intertext model.

5.1.2 An integrated model of epistemic beliefs and documents model representation

In their "integrated model of epistemic beliefs and documents model representation" Bråten, Britt, et al. (2011) propose that beliefs concerning the simplicity of knowledge and beliefs concerning the justification for knowing primarily play a role in the construction of a situations model, whereas beliefs concerning the certainty of knowledge and the source of knowledge are mainly related to the construction of an intertext model. Thus, according to this framework in particular epistemic beliefs concerning the certainty of knowledge and the source of knowledge are related to the evaluation and comparison of multiple information sources and to the interpretation of information in the light of its source.

Bråten, Britt, et al. (2011) further propose that beliefs concerning the certainty of knowledge play a particularly important role when readers are told to construct arguments from what they read, specifically, to express and justify their own opinion or to make informed decisions on a particular issue based on the contents of multiple conflicting texts. Readers who believe that knowledge is tentative and evolving are considered to explore diverse information sources in order to understand different perspectives, to assess authors' expertise and motives (i.e., "who said what for what reason"), and to pay attention to uncertainties. In contrast, readers believing that knowledge is absolute and unchanging are considered to simply try to find one "correct" answer or solution to the problem at hand. As soon as an approved solution to the issue is found, they are likely to terminate the task without paying much attention to the source of the information.

With regard to beliefs concerning the source of knowledge, contrary to the traditional assumption considering the belief that knowledge is transmitted by authorities as being naïve (cf. Hofer & Pintrich, 1997), Bråten, Britt, et al. (2011) suggest this belief to be adaptive when it comes to learning about complex, relatively unfamiliar scientific topics from multiple texts (Bråten et al. 2008; Strømsø et al. 2008). Readers viewing knowledge to be transmitted by authorities are considered to gather information from authorative, and thus, presumably trustworthy information sources, at the expense of discredited or strongly biased sources. In contrast, readers who strongly or one-sidedly believe in the self as a constructor of knowledge, are considered to pay less attention to source characteristics. Instead, they are considered likely to judge the content of the information on the basis of their own opinion, thereby failing to distinguish between more or less trustworthy sources.

To conclude, the integrated model of epistemic beliefs and multiple documents representation by Bråten, Britt, et al. (2011) postulates that not only prior domain knowledge (cf. Chapter 2.3), but also epistemic beliefs about the certainty and source of knowledge play an important role in source evaluations when working with multiple documents. In line with this claim, the next sections provides empirical findings about the relationship between epistemic beliefs and source evaluations during Web search on controversial topics.

5.1.3 Empirical findings regarding the role of personal epistemology in evaluation processes during Web search

Although research on the influence of epistemic beliefs on Web search behavior is still in its infancy, there is growing empirical evidence about the importance of epistemic beliefs for source evaluations during Web search on science-related topics, as will be outlined in following.

Results of a study by Tu et al. (2008) investigating 8th graders performing simple fact-finding tasks as well as complex search tasks on the Web, showed that in complex search tasks (e.g., about the pros and cons of nuclear energy) students with beliefs that knowledge is actively constructed and derived by reason wrote sounder and richer summaries about the search topic than students with beliefs that knowledge is an accumulation of certain and unchanging, isolated facts. In contrast, in simple fact-finding tasks no influences of epistemic beliefs were found, thus corroborating results from hypertext learning on the particular importance of epistemic beliefs in the context of complex ill-structured problems (cf. Schraw et al., 1995).

Other studies investigated the effects of epistemic beliefs on searchers' ongoing evaluation processes during complex Web search tasks through the use of different trace-methodologies. Mason and Ariasi (2010), for example, showed by using eye-tracking methodology that depending on their domain-specific epistemic beliefs university students visually attended to different parts of a Web page and visually inspected different kind of Web pages with varying intensity during their Web search on a science-related topic in the domain of molecular biology. Whereas the epistemic conception that knowledge is certain and unchanging led to spending more time reading the most well-known information source, the conception that knowledge is tentative and evolving led to concentrating more on controversially discussed, newer information, given that it was provided by an authoritative information source. Similarly, Kammerer, Wollny, Gerjets, and Scheiter (2009) found for a complex Web search task about the "low carb versus low fat controversy" (cf. Study 1), that university students' general epistemic beliefs influenced their attention allocation on search results presented by a search engine, depending on the interface design of the SERPs (see chapter 5.2 for more details).

Furthermore, the results of a thinking-aloud study by Whitmire (2004) on undergraduate students' Web search behavior regarding two controversial topics (i.e., the man-made construction of the pyramids, food safety) showed that students' domain-general epistemic beliefs affected their source evaluations. Similar to empirical results found for students with high prior domain knowledge (e.g., MaKinster et al., 2002), students who viewed knowledge as more tentative and complex examined the URLs of search results and paid attention to author and affiliation information on Web pages in order to evaluate information quality during Web search. Moreover, they did not reject conflicting information during document collection and they were capable to differentiate between authoritative and non-authorative information sources. In contrast, students who tended to believe that knowledge is certain or absolute preferred information sources that supported their views on the topic over information sources that were in opposition to their views. When they encountered conflicting information sources, instead of evaluating the trustworthiness of these sources by themselves, they preferred to ask authority persons (e.g., their professor) about the trustworthiness of the sources. Furthermore, they tended to base their selection decisions on the title and the page excerpt of the search result or on the ranking position because they believed that the search engine placed the best search results first.

In another recent retrospective interview study, Mason, Boldrin, and Ariasi (2010a) investigated the relationship between middle-school students' domain-specific beliefs about science and their epistemic reflections about evaluation processes in the context of a complex and controversial Web search task (about the extinction of the dinosaurs). Verbal data obtained in retrospective interviews revealed that the more students believed in the complexity of scientific knowledge, the more they reflected on scientific evidence provided by the information sources and the more they critically compared information from multiple information sources. Furthermore, the more advanced students' beliefs about science, the better were there learning outcomes about the search topic at hand. In this study, however, students were explicitly asked about their quality-related evaluation strategies.

Instead of using questionnaires to assess epistemic beliefs, in a series of thinking-aloud and interview studies Hofer (2004) examined epistemic reflections of high-school and college students who searched the Web for information on a scientific topic (the communication behavior of bees). Findings indicated that students who

considered knowledge to consist of certain and unchanging, isolated facts conducted the search task in a brief and perfunctory way, without pursuing additional information sources or reflecting on the credibility and accuracy of the information sources they inspected. In contrast, students' epistemic awareness about the tentative and evolving nature of knowledge, which was related to a higher expertise in science, led students to actively seek for more recent sources and to pursue informed strategies for searching.

Whereas the previously reported studies examined users' domain-general or domain-specific epistemic beliefs, as mentioned above Bråten et al. (2005) and Strømsø and Bråten (2010) assessed university students' Internet-specific epistemic beliefs using the ISEQ. In line with Hofer and Pintrich's (1997) conceptualization of epistemic beliefs, the ISEQ was designed to measure the following four belief dimensions: (1) beliefs about the *certainty* of Web-based knowledge (e.g., "On the Internet many different sources provide the correct answer to questions related to my course work"), (2) beliefs about the *structure* (i.e., *simplicity*) of Web-based knowledge (e.g., "The strength of the Internet is the vast amount of detailed information that is located there about what I am studying"), (3) beliefs about the Web as a *source* of knowledge (e.g., "I am most confident that I have understood something for my classes when I have used the Internet as a source"), and (4) beliefs about the *justification* of knowledge claims encountered on the Web (e.g., "To check whether the study-related knowledge I find on the Internet is reliable, I try to evaluate it in relation to other knowledge I have about the topic"). Based on the results of a factor analysis, however, certainty beliefs and source beliefs were merged into one dimension (cf. Strømsø & Bråten, 2010).

The items of the remaining three scales differ with regard to how unequivocally their meaning can be interpreted in terms of epistemic sophistication. With regard to beliefs concerning the justification of knowledge hardly anyone with a certain level of epistemic awareness would hesitate to agree that Web-based knowledge claims need to be checked against other sources, reason, and prior knowledge (cf. Strømsø & Bråten, 2010). However, regarding the other two scales the respective ISEQ items are somewhat ambiguous with regard to whether they mean that a) *all* or at least *most* knowledge on the Web has particular features (e.g., to be correct), or that b) at least instances of such kind of knowledge *can* be found on the Web, with the latter being a

more appropriate belief than the first. This should be considered, when interpreting the study results presented in the following.

In two questionnaire-based studies the Internet-specific epistemic beliefs of political science students (Bråten et al., 2005) or physics students (Strømsø & Bråten, 2010), respectively were correlated with students' self-reports on their Web search behavior when searching for study-related information. Results by Strømsø and Bråten (2010) indicated a positive relationship between the belief that Web-based knowledge claims need to be checked against other sources, reason, and prior knowledge and the reported use of self-regulatory strategies (such as planning, monitoring, and regulating cognition and performance) during Web search on study-related information. Moreover, results of both studies indicated that also beliefs that the Web is a reliable resource that contains correct and detailed facts were positively related to the reported use of self-regulatory strategies. Furthermore, undergraduate students with such beliefs reported to be more competent and to experience less problems in conducting Web searches, evaluating information sources, and using the found information for their course-work than students who had doubts about the Web containing correct and detailed facts (Bråten et al., 2005; Strømsø & Bråten, 2010).

Bråten et al. (2005) and Strømsø and Bråten (2010) point out two alternative explanations for the findings with regard to the beliefs about whether or not the Web contains correct and detailed facts. On the one hand, it can be considered veridical that the Web actually does contain a wealth of detailed and factual, correct information. Thus, students possessing respective beliefs might be those who have developed skills to find high-quality information on the Web and to differentiate it from low-quality information. On the other hand, however, it is also conceivable that students who believe that the Web is a reliable resource that contains correct and detailed facts may be less likely to realize the great challenge involved in managing the wealth and heterogeneity of information sources available on the Web (cf. Bråten et al., 2005). Hence, these students might be treating Web search as a relatively unproblematic activity, without being aware of any risks related to the use of biased or inaccurate information. As the results are only based on self reports, which may be prone to social desirability biases, it is possible that students' self reports contradicted their spontaneous search and evaluation behavior during Web search. Indications for this assumption are given by Flanagin and Metzger (2007) who found a negative relationship between self-reported and observed evaluation strategies to verify the

information provided on a Web page. Furthermore, university students predominantly acquire scientific, study-related knowledge through quality-controlled resources such as lectures, textbooks, and scientific journals rather than from the Web (although some of these materials can be downloaded from the Web). Thus, epistemic doubts with regard to the quality of study-related knowledge available on the Web as compared to respective knowledge provided by other resources seem appropriate.

Based on these considerations, in the present dissertation students with strong beliefs that the Web contains correct knowledge were assumed to be less aware about the necessity to critically evaluate the information found on the Web than those students who have doubts about that issue. Yet, in order to test the assumption that epistemic doubts that the Web contains correct and detailed knowledge facilitate the evaluation of information quality during Web search, more direct measures of actual Web search behavior, such as log files and eye-tracking data (cf. Study 1 of this dissertation), are needed. This was one of the central aims of the two studies presented in the remainder of this dissertation.

5.2 The role of the search interface in evaluating information quality

Even if users are epistemically aware of the varying information quality on the Web, the activation of their epistemic beliefs in the context of a particular Web search task might depend on the search interface (e.g., the SERP) they have at their disposal to evaluate and select potentially relevant Web pages. Thus, the concrete enactment of quality-related evaluation processes during Web search might depend on whether the search interface affords them. The interface of popular search engines such as Google, however, has two main characteristics which rather hamper than facilitate the evaluation of information quality during search result selection. First, search engines usually present search results in a vertical list, with the most topically relevant and most popular Web pages being the highest-ranked ones (cf. Cho & Roy, 2004). Such a rank-ordered list format suggests to start reading at the top of the list and to follow to the strict and non-ambiguous order when reading and selecting search results (O'Brien & Keane, 2006). Second, search engines usually display only a few lines of information for each search result (e.g., a title, an excerpt from the

respective Web page, a URL) on which evaluation processes aimed at deciding which search results to select for further inspection must be based. Moreover, the search result descriptions are typically mostly confined to content information, with the search terms used for the search highlighted in bold in the search results. In contrast, quality-related source information in the search results descriptions is sparse and non-salient (e.g., hidden in the cryptic URLs of the search results). Thus, in particular for searchers with low prior domain knowledge standard SERPs might not provide sufficient affordances to activate their epistemic beliefs. Accordingly, as has been shown in Study 1 as well as in previous studies (e.g., MaKinster et al., 2002) instead of taking over the responsibility for the cognitively demanding task of evaluating and selecting search results on their own, low-knowledge searchers tend to rely on superficial, but salient cues such as the ranking position of the search results or the keywords displayed in the search results.

In sum, it can be expected that (1) the salience of the ranking (or ordering) of search results as well as (2) the insalience of proximal cues in search result descriptions pointing to the quality of information are major constraints of spontaneous source evaluations for low-knowledge searchers. On the contrary, a search interface that refrains from making the ranking of search results the most salient feature might stimulate searchers who are aware about the varying information quality on the Web to engage in source evaluations to decide which Web pages to access for further inspection. Furthermore, a search interface that not only refrains from making the ranking of search results the most salient feature, but also provides salient proximal cues that point to information quality might compensate for a low epistemic awareness of the varying information quality on the Web. To conclude, within the limits of users' individual cognitive and metacognitive prerequisites, a proper search interface might lead to navigational decisions that are based to a substantial degree on evaluations of information quality.

The following two sections present theoretical considerations as well as empirical findings concerning the impact of the (in)salience of the ranking or ordering of items (5.2.1) and of the (in)salience of quality-related source cues (5.2.2) on subjective ratings, navigation and viewing behavior, and evaluation processes.

5.2.1 Salience of the ranking or ordering of items: Lists versus alternative presentation formats

The influence of presentation order of items presented in a vertical list on subjective rank ratings of the respective items has already been shown in 1936 by Pervical Symonds. In Symond's study high school students were required to rank 15 printed items about problems and interests of adolescents with regard to importance, with the items being presented in a vertical list in two reversed orders. Results showed that items presented in the top of the list were over-ranked and items on the bottom of the list were under-ranked by the students.

To investigate whether such a "fondness for items at the beginning of […] lists" (Keane, O'Brien, & Smyth, 2008, p.51) also holds for SERP lists, Keane et al. (2008) and Pan et al. (2007) used a methodological paradigm similar to the one by Symonds (1936). They experimentally manipulated the relevance order of ten search results on a Google SERP by presenting search results either in a regular order, or in a systematically reversed 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 who were required to conduct fact-finding tasks visually inspected more search results than when the top search results were the most relevant ones. As a consequence, 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, even when they were the least relevant results.

Furthermore, O'Brien and Keane (2006) found equivalent effects with a search result list displayed as a plain text interface (without any Google logo, etc.), contradicting an explanation given by Pan et al. (2007) that the influence of the ranking only arises from people's implicit trust in the search engine rather than from the fact the search results are presented in a list. A study by Guan and Cutrell (2007) that varied the position of the target search result (i.e., the most relevant result) to accomplish a fact-finding task 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 visually inspected more search results than when the target search result was among the top positions on the SERP. In sum, these empirical findings indicate that Web users searching for specific facts are heavily influenced by the ranking of

the search engine. Although searchers seem to employ some degree of evaluations regarding the topical relevance of the search results presented on a SERP, they often tend to neglect their own evaluations in favor of obeying the ranking determined by the search engine.

Furthermore, as has been shown in Study 1 of this dissertation, also in a more complex Web search task users spent more attention to the top half of the search results on the SERP than to the bottom half. Moreover, in particular searchers with low prior domain knowledge selected the top half of the results more often than the bottom half. MaKinster et al. (2002) reported on similar results with low-knowledge students simply selecting the search results in the order in that they were presented by the search engine. Whereas the previously reported studies only examined users' selection decisions in list interfaces, two other experimental studies outlined in the following directly compared participants' viewing and navigation patterns when working with a list format or with an alternative presentation format.

Salmerón, Gil, et al. (2010) investigated university students' navigation behavior during Web search on the topic of climate change. To complete the Web search task participants either received search results provided in a standard Google-like list interface or in a graphical-overview interface similar to the search engine *Kartoo* (closed down in January 2010) in which search results were displayed by means of a graphical overview indicating the semantic relationships between the search results. Navigation results revealed that in the graphical overview interface students' navigation patterns were significantly more heterogeneous than in the list interface (as measured by the Levenshtein distance, see Chapter 6.2.5.1 for details). Whereas in the list interface students adhered to a linear top-to-bottom navigation pattern, the graphical overview interface supported a more free selection sequence of the available search results. In addition, the exploration of the Web pages by using the graphical overview interface resulted in better integration of information from different Web pages (i.e., resulted in a stronger intertext model) than when working with the standard list interface.

In a newspaper reading study Holsanova, Holmberg, and Holmqvist (2008) compared participants' reading paths in an information graphic (i.e., a graphical visualization consisting of a series of text paragraphs and pictures) that was either presented in a serial format or in a radial format. Whereas the serial information

graphic had a sequentially organized format suggesting a specific reading path, the radial information graphic did not indicate any specific reading path nor a certain entry point where to start reading the information graphic. Results from eye tracking data showed that in contrast to the serial information graphic in which readers homogeneously followed the suggested reading path in an almost linear order, in the radial information graphic readers showed high variations with regard to their reading sequences.

To conclude, the studies reported in this section provide evidence that list formats set a strong focus on the items (e.g., search results) presented in the top of the list and are likely to impose a strict and non-ambiguous order in which to read and select items. In contrast, alternative presentation formats have been shown to result in less linear and more heterogeneous viewing and navigation patterns. Accordingly, a search interface that reduces the salience of the ranking by presenting search results in a format different from a single, vertical list can be assumed to stimulate searchers to explore the search results more freely, and to base their selection decisions less on the position than a standard list format. Searchers who are aware about the varying information quality on the Web thus might base their selection decisions more on own evaluations of the search results, for instance, in terms of evaluating the credibility of the information sources. This design approach of an alternative SERP format with a reduced salience of the ranking was examined in Study 2 of this dissertation (Chapter 6).

5.2.2 Salience of quality-related source information

Given the fact that link descriptions used in standard Web search or hypertext environments are mostly confined to content information, a few studies investigated the effects of adding specific source cues to the content information of the respective link descriptions, yielding positive results with regard to evaluation and navigation behavior.

Ivory, Yu, and Gronemyer (2004) enhanced Google SERPs with additional quality-related cues for each search result indicating the number of graphical ads, the number of words, and the estimated quality of the corresponding Web page. They used simple fact-finding tasks to demonstrate that adding this type of quality-related cues improved participants' ability to select search results that led to the correct answer

and to reject inappropriate search results. For more complex information search tasks, which do not possess one correct solution, quality-related cues, however, might have an even greater impact.

This assumption is supported, for example, by findings from Schwarz and Morris (2011) who augmented search results with visualizations that represented, for instance, the page rank of the Web page, its popularity among domain experts, and a Web page's receipt of awards or certifications. In their experimental study, for which they chose complex search topics from the field of politics or health care, participants were asked to rate the credibility of Web pages on the basis of search results that were presented to them one at a time. Results showed that the augmented search results improved users' credibility evaluations of Web pages as compared to standard search results. In addition, survey data indicated that participants found the information about the popularity of the Web page among experts as most helpful for credibility evaluations.

Winter, Krämer, Appel, and Schielke (2010) investigated Web users' credibility evaluations when searching information on the topic of "violence in the media" by using an experimentally designed Web platform that provided a list of article descriptions (similar to search results). In this platform, in addition to a title and short description of the articles, the name and profession of the authors as well as community ratings about the expertise of the authors were displayed for each "search result". Results showed that articles provided by domain experts (operationalized via the profession of the authors) were selected more often and were rated more credible than articles provided by non-experts. In contrast, community ratings about the expertise of the author represented by rating stars did not have a significant effect on individuals' credibility rating and information selection. However, if only non-expert authors were taken into account, articles of authors with a high community rating tended to be selected more often than articles with a low rating.

With regard to the online news service Google News, a study by Sundar, Knobloch-Westerwick, and Hastall (2007) showed that the availability of quality-related cues (i.e., information on the source, the recency of a story, and the number of related articles) for each news item affected users' subjective evaluations of the news leads. Similar to the results by Winter et al. (2010) the source information had the strongest effects on users' credibility judgments regarding the news leads, with news from

credible sources being preferred over news from sources of rather low credibility. Furthermore, when the source was judged as being of low credibility, the recency of the story and the number of related articles also had a strong impact on users' judgments regarding the credibility of the news lead.Whereas the previously reported studies used rating scales (or a combination of rating scales and selection data) to examine users' credibility evaluations, which might have distorted participants' spontaneous evaluation processes (cf. Chapter 3.4), Kammerer et al. (2009) investigated university students' spontaneous Web search behavior by means of eye-tracking and log file data. In their experimental study participants either used a standard Google search result list or an augmented search result list that additionally contained source category labels printed in bold next to the URL (similar to the search catalogue Clewwa, http://www.clewwa.de/) to conduct a complex Web search task about the "low carb versus low fat controversy" (cf. Study 1). The five source categories were "Science/Institutions", "Portals/Advisors", "Journalism/TV", "Readers' Comments", and "Shops/Companies" (translated from German). Results showed that the availability of source categories influenced students' viewing- and navigation behavior. For instance, on the augmented SERPs participants showed less linear viewing sequences than on standard SERPs. Furthermore, participants who used the augmented SERPs spent more time on scientific or official Web pages, but less time on television or journalistic Web pages during Web search than participants who accessed Web pages via standard SERPs (Kammerer, Wollny, & Gerjets, 2011). Beyond that, as compared to the standard SERPs, the availability of source categories on SERPs caused students with epistemic beliefs that knowledge is certain and provided by omniscient authorities to fixate less on the search results belonging to all but the "Science/Institutions" category, indicating that they made their selections based on the source categories provided. In contrast, students with beliefs that knowledge is uncertain and derived by reason did not differ in their total fixation time on search results in the two experimental conditions.

In sum, the reported studies provide evidence that the availability of salient quality-related source cues in search results, article descriptions, or news items stimulates users' source evaluations during information seeking, resulting in an increased selection of high-quality, authoritative, and trustworthy information sources than when such quality-related cues are absent. Moreover, the findings by Kammerer et

al. (2009) indicate that the provision of quality-related source cues might compensate for a low epistemic awareness of the varying information quality on the Web.

Instead of providing source information for each search result, article description, news item, or hyperlink separately as in the previously reported studies, another approach is to group search results according to categories. A study by Dumais, Cutrell, and Chen (2001), for example, revealed that interfaces with search results being spatially grouped according to categories (in their case, semantic categories), resulted in a more efficient selection behavior than list interfaces in which the same category labels were provided for each individual search result. This effect might be explained by the concept of representational guidance from instructional psychology (e.g., Suthers & Hundhausen, 2003), that providing users with categories helps "to structure the task domain […] [and] might guide information search" (Stadtler & Bromme, 2008, p. 220). Furthermore, the categories might help searchers to internalize and eventually think in terms of the categories (Stadtler & Bromme, 2008).

Accordingly, a search interface that provides additional source cues and a grouping of search results according to these source categories can be assumed to better guide low-knowledge Web searchers, in particular those with a low epistemic awareness of the varying information quality on the Web, in their search and to support their evaluations of information quality than a standard list format. This design approach was examined in Study 3 of this dissertation (Chapter 7).

5.3   Summary and overview of Study 2 and Study 3

The preceding paragraphs have illustrated the role of Web users' epistemic beliefs as an individual variable and of the search interface as a resource variable (as classified according to the conceptual framework of Lazonder and Rouet, 2008, see Chapter 2.3) in information seeking and evaluation activities, when working with multiple printed or Web-based documents. According to the existing evidence, these two variables can be considered to play a crucial role for the evaluation of information quality on SERPs. Accordingly, appropriate epistemic beliefs and a proper search interface might compensate for low prior domain knowledge and thus might result in a critical evaluation of information quality during search results selection.

Based on amounting evidence that a person's epistemic beliefs about the nature of knowledge and knowing affect the comprehension of multiple texts, Bråten, Britt, et al. (2011) proposed an integrated model of epistemic beliefs and documents model representation, specifying how and why epistemic beliefs might facilitate or impair the construction of a documents model. As outlined in section 5.1.1, with regard to source evaluations as part of the formation of an intertext model they identified beliefs concerning the certainty of knowledge and the source of knowledge to play a particularly important role. Furthermore, there is growing empirical evidence from research using trace-methodologies such as eye tracking or thinking aloud that epistemic beliefs play an important role in source evaluations during Web search on controversial topics (e.g., Hofer, 2004; Mason et al., 2010; Whitmire, 2004). Epistemic beliefs that knowledge is complex, interconnected, tentative and evolving, as compared to epistemic beliefs that knowledge consists of certain and unchanging, isolated facts, seem to stimulate Web-based source evaluations. Yet, with regard to Internet-specific epistemic beliefs (Bråten et al., 2005; Strømsø & Bråten, 2010) the current state of the art allows no clear conclusion whether beliefs that the Web is a reliable resource that contains correct and detailed facts hamper quality-related evaluation processes during Web search or rather facilitate them. Therefore, a central aim of the experimental studies described in the following two chapters was to shed light on the relationship between Internet-specific epistemic beliefs and source evaluations on SERPs during Web search on a complex medical issue. Based on the theoretical considerations by Bråten, Britt, et al. (2011), that in particular the two dimensions certainty of knowledge and source of knowledge play a crucial role for source evaluations when working with multiple documents, the scale about "certainty and source of knowledge" of the ISEQ (Strømsø & Bråten, 2010) was used as epistemic beliefs variable in the present dissertation.

However, as discussed in the previous chapter whether users' epistemic beliefs are activated during Web search might depend on the search interface they have at their disposal to decide which Web pages to access for further inspection. The typical SERP interface that (a) presents search results in a vertical list imposing a high salience on the ranking of the search results and that (b) provides only sparse and non-salient quality-related source information, does not seem to sufficiently stimulate or support the evaluation of information quality during Web search. Instead such a list format has been shown to set a strong focus on the search results presented

in the top of the list and to impose a strict and non-ambiguous order in which to read and select the presented search results (e.g., Keane et al., 2008; Pan et al., 2007; Salmerón, Gil, et al., 2010). In contrast, previous research indicates that alternative presentation formats that provide salient proximal cues for information quality and/or display search results in a format different from a list might stimulate searchers to engage in quality-related evaluation processes on SERPs (e.g., Kammerer et al., 2009; Salmerón, Gil, et al., 2010; Schwarz & Morris, 2011).

Considering both the role of epistemic beliefs and of the search interface, a search interface that refrains from making the ranking of search results the most salient feature might particularly stimulate searchers who possess a certain degree of awareness about the varying information quality on the Web to engage in source evaluations. In contrast, a search interface that not only refrains from making the ranking of search results the most salient feature, but also provides salient proximal cues that point to information quality might have the potential to compensate for a low epistemic awareness of the varying information quality on the Web (cf. Kammerer et al., 2009). Thus, such a search interface might particularly help searchers with less appropriate epistemic beliefs to access high-quality information sources. Based on these assumptions in the two experimental studies presented in Chapters 6 and 7 of this dissertation two different interface design approaches of SERP layouts were implemented to test whether they stimulated or supported searchers with certain levels of epistemic awareness to engage in source evaluations on SERPs: Whereas the first interface design approach – examined in Study 2 – focused on reducing the salience of the ranking of search results by presenting the search results in a grid interface (i.e., a three-by-three grid), the second design approach – investigated in Study 3 – aimed at both reducing the salience of the ranking and simultaneously increasing the salience of quality-related source information in the SERPs by presenting search results in a tabular interface with labeled columns, in which search results are grouped according to specific source categories.

As in Study 1, to examine searchers' evaluation processes on SERPs Study 2 and 3 used eye tracking methodology, log file data (i.e., mouse clicks), and verbal protocols as process measures. In addition, the product of the search task conducted by the participants was analyzed by asking them to list arguments (Study 2) or to write an argumentative essay (Study 3) about the search topic.

Because the goal of these studies was to examine whether there exist preconditions under which even searchers with low prior domain knowledge engage in source evaluations on SERPs, prior domain knowledge was kept constantly low by choosing an unfamiliar search topic: The search task used in Study 2 and 3 was to seek information on the WWW about two competing, controversially discussed therapies for Bechterew's disease, that is, a rare chronic rheumatic disease affecting the spine. Besides, contextual variables were held constant as well, not providing any evaluation instructions before or during the Web search.

6   Study 2: The role of Internet-specific epistemic beliefs and of the salience of the ranking in search interfaces in evaluating information quality during Web search on medical and health-related issues

6.1   Research questions and hypotheses

As has been shown in Study 1 of this dissertation, during Web search on a controversial health-related issue in particular students with no or little prior domain knowledge only rarely evaluated the quality of information on SERPs and on Web pages and to a great extent simply relied on the ranking provided by the search engine. Furthermore, whereas explicit evaluation instructions enhanced quality-related evaluation processes of students with a certain amount of prior domain knowledge, such instructions had no effects on the source evaluations of low-knowledge students.

Therefore, the goal of the present study was to examine other factors, namely appropriate Internet-specific epistemic beliefs (individual variable) and affordances provided by the search interface (resource variable), that might facilitate source evaluations when searching on the Web for complex medical and health care topics, for which prior knowledge is low. In particular, the focus was on the analysis of low-knowledge searchers' initial source evaluations on SERPs, that is, on their decisions with regard to which Web pages to access for further inspection.

As outlined in Chapter 5.1 both theoretical considerations (Bråten, Britt, et al., 2011) as well as empirical findings (e.g., Hofer, 2004; Mason et al.,2010; Tu et al., 2008; Whitmire, 2004) indicate that epistemic beliefs play an important role in making evaluations of information quality during Web search or during studying multiple print documents. With regard to Internet-specific epistemic beliefs, Bråten et al. (2005) suggest that students with high beliefs that the Web is a reliable resource that contains correct knowledge might not realize the epistemic challenges involved in managing the wealth of information and evaluating the different types of information

sources. Thus, they might be rather unaware of the necessity to engage in source evaluations during Web search. In line with this assumption, in the present study doubts (i.e., low beliefs) that the Web contains correct knowledge about scientific, study-related contents (as measured by the "certainty and source of knowledge" scale of the ISEQ by Strømsø and Bråten, 2010) were considered to be an important precondition to critically evaluate the quality of Web information. However, whether such epistemic doubts with regard to the quality of Web information are activated in the context of a certain Web search task might depend on the search interface Web users have at their disposal to evaluate and select potentially relevant Web pages.

As outlined in Chapter 5.2.1, a search interface that presents the search results returned by the search engine in a ranked-ordered vertical list (i.e., a high salience of the ranking of search results) might not sufficiently activate epistemic doubts with regard to the quality of Web information. Rather, such a standard SERP format suggests to simply rely on the ranking position of the search results (cf. Pan et al., 2007); in particular when cognitive resources for source evaluations due to a low prior domain knowledge are limited (MaKinster et al., 2002). In contrast, a search interface that refrains from making the ranking of search results the most salient feature might be more likely to activate epistemic doubts with regard to the quality of Web information. To test this assumptions, in the present experimental study participants either received search results in a Google-like *list interface* with search results presented in a rank-ordered list or in a *grid interface* with search results presented in a three-by-three grid, similar to the search engine *Viewzi* (see Figure 17), which was closed down in January 2011. This organization implies that there is no strict and non-ambiguous order which can be obeyed when reading and selecting search results. In particular, it remains unclear for the searcher 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, according to a list structure), or whether there exists a ranking at all. Therefore, in a grid interface the decision about the reading and selection order is left open to the user. The fact that the ranking is less salient might encourage users to engage in own evaluations that might help to decide which search results to select.
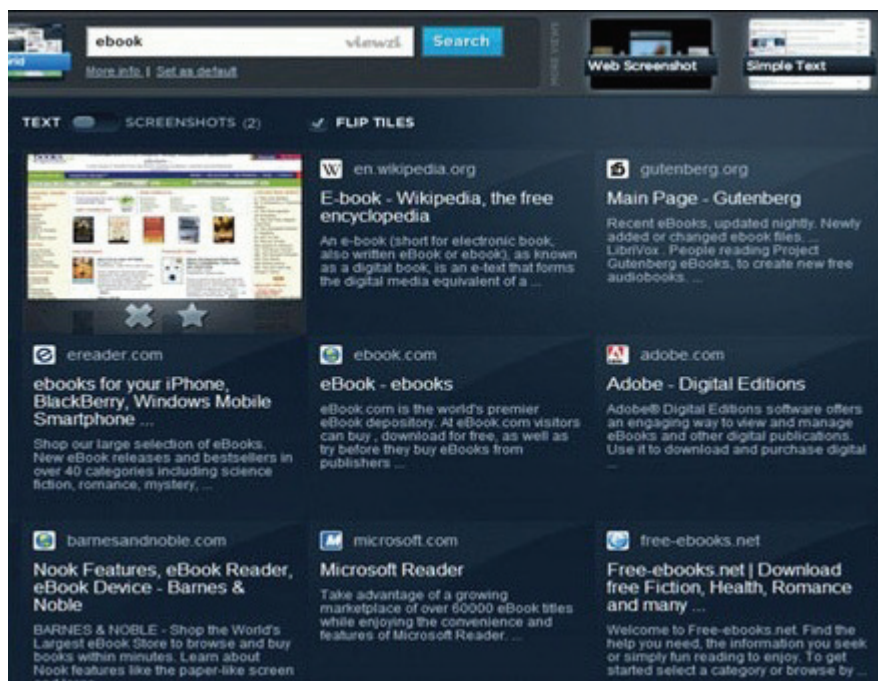
Figure 17. Screenshot of the search engine *Viewzi* (closed down in January 2011) that presented search results in a grid format.

It was hypothesized that the grid interface as compared to the list interface would reduce the focus on the top positions, and thereby would stimulate users to evaluate search results according to other criteria, including information quality. Thus, the use of the grid interface during Web search should also result in higher-quality search outcomes than the list interface. However, because people with high beliefs that the Web contains correct knowledge might still see no need to evaluate the quality of Web information, the advantage of the grid interface over the list interface on participants' engagement in source evaluations was assumed to be *larger* for people with doubts (i.e., low beliefs) that the Web contains correct knowledge. Figure 18 graphically represents this assumed interaction between search interface and epistemic beliefs on participants' engagement in source evaluations during Web search.
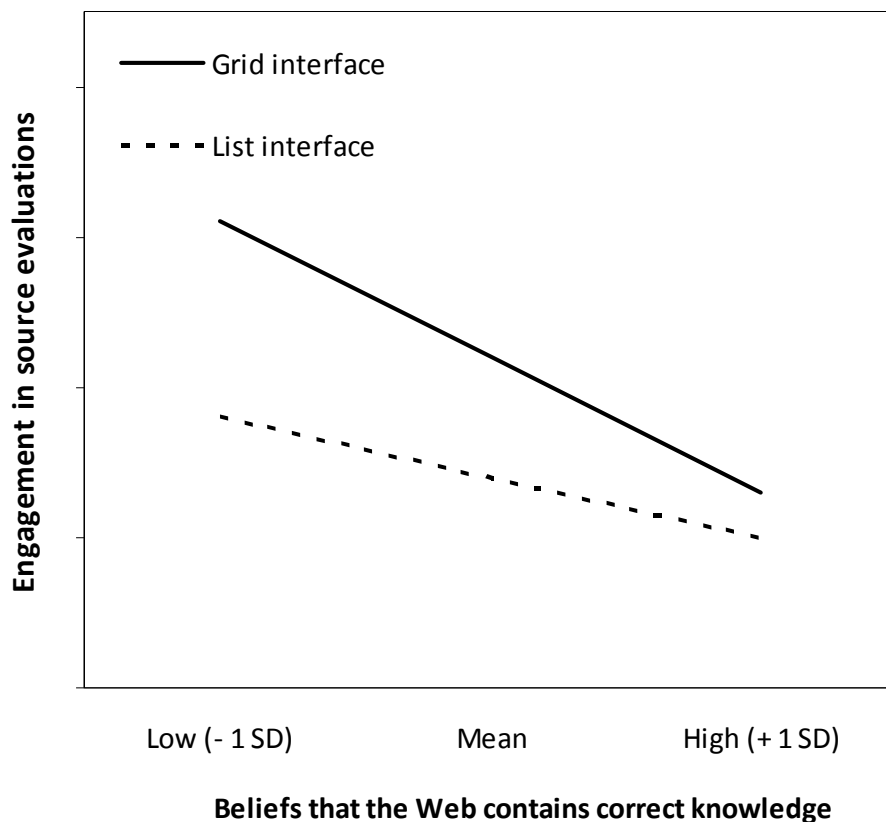
Figure 18. Assumed interaction pattern between search interface and epistemic beliefs on the engagement in source evaluations.

To investigate the extent to which participants engaged in source evaluations on SERPs, a methodological paradigm similar to the one by Keane et al. (2008) and Pan et al. (2007), which was outlined in Chapter 5.2.1, was used. Whereas these authors had experimentally manipulated the relevance order of the search results on a SERP, in the present study the "trustworthiness order" of the search results was manipulated. As described in Chapter 2.2 trustworthiness refers to the perceived willingness of an information source to provide accurate and unbiased information (cf. Danielson, 2005), thus constituting an important criterion of the multidimensional construct of information quality. Accordingly, in the present study search results were presented either in an *optimal trustworthiness order* with the search result rated as most trustworthy being on top of the list or in the upper left corner of the grid, respectively, or in a *reversed trustworthiness order* with the search result rated as most trustworthy being at the bottom of the list or in the bottom right

corner of the grid, respectively[2]. The latter, reversed trustworthiness order represented the critical condition in which simply relying on the ranking order would result in the selection of search results that were of rather low trustworthiness according to ratings obtained in a pilot study where 24 university students rank-ordered the search results according to their perceived trustworthiness (see section 6.2.2. for details). Accordingly, in the materials used in this study search results of low trustworthiness were predominantly those that linked to commercial Web sites or to contributions in blogs or forums.

Evaluation processes on SERPs in the present experimental study were assessed through a combination of eye tracking methodology, log file data (i.e., mouse clicks), and verbal protocols. In addition to process measures, the quality of the solution to the information problem was analyzed by asking participants to list arguments with regard to the conflicting medical information problem.

According to these measures, the assumption that both a search interface with a reduced salience of the ranking of search results and a high epistemic awareness of the varying information quality on the Web would facilitate source evaluations on SERPs, with the combination of both having the strongest effects, resulted in the following more specific hypotheses.

6.1.1 Eye-tracking data – Viewing sequences

Due to the reduced salience of the ranking, the grid interface was expected to cause a more free exploration of the search results than the list interface (cf. Salmerón, Gil, et al., 2010). Thus, it was hypothesized that in the grid interface the search results presented on a SERP would be visually inspected in a less linear sequence and with higher variations across individuals than in a list interface. In the list interface, instead, in line with results from previous studies (e.g., Pan et al., 2007) participants were assumed to homogeneously show rather linear viewing sequences from top to bottom when inspecting the search results available on a SERP (Hypothesis 1). Epistemic beliefs were not expected to affect participants' viewing sequences.

---

[2] For the grid interface search results were arranged line-by-line, that is, from left to right in each of three rows. However, this was unknown to the participants (and is also unknown to users of the *Viewzi* search engine).

6.1.2 Eye-tracking data – Attention to search results

With an increased fixation time on search results being an indicator for more thorough evaluation processes (cf. Study 1), it was assumed that the more participants doubted that the Web contains correct knowledge, the longer they would attend to search results on the SERPs (Hypothesis 2a). Furthermore, in line with findings from Study 1, it was assumed that in the list interface more attention would be given to the top search results than to the bottom search results, irrespective of the trustworthiness order. In contrast, in the grid interface because of the uncertainty about the ranking of the search results, when the top search results were of low trustworthiness (reversed trustworthiness order) participants were expected to not only attend to the top search results but also to the rest of the search results presented on the SERP in order to check whether those were "better", that is, more trustworthy than the top results. Therefore, it was hypothesized that when the top search results were of low trustworthiness (reversed trustworthiness order) grid interface users would attend shorter to those least trustworthy search results and longer to the most trustworthy search results than list interface users. In the optimal trustworthiness order no differences were expected between the list interface and the grid interface with regard to their attention to search results, resulting in an interaction effect between interface and trustworthiness order (Hypothesis 2b). Furthermore, the advantage of the grid interface over the list interface in the reversed trustworthiness order with regard to the attention to trustworthy search results was expected to be larger for participants with doubts (i.e., low beliefs) that the Web contains correct knowledge than for participants with high beliefs about this issue. (Hypothesis 2c).

6.1.3   Search result selection

In line with findings from Study 1 about low-knowledge students' selection behavior, it was assumed that in the list interface due to the high salience of the ranking the higher-ranked search results would be selected more often than the lower-ranked search results, irrespective of the trustworthiness order. In contrast, it was assumed that the grid interface because of the uncertainty about the ranking of the search results would stimulate participants to evaluate the search results on their own, also in terms of information quality. Therefore, it was hypothesized that when the top search results were of low trustworthiness (reversed trustworthiness order) in the grid interface the most trustworthy search results would be selected more often and the

least trustworthy search results less often than in the list interface. In the optimal trustworthiness order no differences were expected between the list interface and the grid interface with regard to the selection of search results, resulting in an interaction effect between interface and trustworthiness order (Hypothesis 3a). Furthermore, the advantage of the grid interface over the list interface in the reversed trustworthiness order with regard to the selection of trustworthy search results was expected to be larger for participants with doubts (i.e., low beliefs) that the Web contains correct knowledge than for participants with high beliefs about this issue (Hypothesis 3b).

6.1.4 Verbal utterances

In line with findings from Study 1 about low-knowledge students' verbal utterances, in the list interface quality-related utterances were expected to be rare. Because of the reduced salience of the ranking in the grid interface, it was hypothesized that the grid interface would increase the number of quality-related verbal utterances (e.g., about the type or credibility of information sources) and, furthermore, would decrease the number of verbal utterances addressing the position of search results in a SERP as compared to the list interface (Hypothesis 4a). Moreover, the increase of quality-related utterances in the grid interface as compared to the list interface was assumed to be larger for participants with doubts (i.e., low beliefs) that the Web contains correct knowledge than for participants with high beliefs about this issue (Hypothesis 4b). In contrast, the decrease of utterances addressing the position of search results was expected to be larger for participants with high beliefs that the Web contains correct knowledge than for participants with doubts about this issue (Hypothesis 4c).

6.1.5 Quality of information problem solving

If in the reversed trustworthiness order grid interface users selected more of the most trustworthy search results than list interface users, then the quality of grid interface users' solution to the information problem (i.e., their search outcome) should improve as well. It was assumed that when the top search results were of low trustworthiness the users of the grid interface would list more arguments from the most trustworthy information sources than list interface users. In the optimal trustworthiness order no differences were expected between the list interface and the grid interface with regard to participants' search outcome, resulting in an interaction

effect between interface and trustworthiness order (Hypothesis 5a). Furthermore, the advantage of the grid interface over the list interface in the reversed trustworthiness order with regard to the quality of the search outcome was expected to be larger for participants with doubts (i.e., low beliefs) that the Web contains correct knowledge than for participants with high beliefs about this issue (Hypothesis 5b).

6.2   Method

6.2.1 Participants

Participants were 80 university students (17 male, 63 female; $M = 25.40$ years, $SD = 3.95$; range 19 – 39 years) from different majors at the University of Tuebingen, Germany; participation was rewarded with either course credit or payment. Pharmacy and medical students were excluded from the study. Prior knowledge on the content domain of the study (i.e., Bechterew's disease and other rheumatic diseases) was assessed (see section 6.2.3.5) to ensure that prior domain knowledge was low. One participant was dropped from all further analyses because of a prior domain knowledge score of 3.60. The rest of the participants had scores from 1.00 to 2.40 ($M = 1.32$, $SD = 0.38$) on a scale from 1 (low) to 5 (high). Participants had normal or corrected to normal vision. All participants reported to use Google as their primary search engine.

6.2.2 Experimental design

The experiment was based on a four-factorial mixed-model design. As a first factor the *search interface* was varied between subjects. Participants either received SERPs in the form of a standard *list interface* (see Figure 19) with nine search results being listed from top to bottom on each SERP, or in the form of a *grid interface* (see Figure 20) with the same nine search results being arranged in a three-by-three grid similar to the search engine *Viewzi*. In the grid interface 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. However, participants were unknown about this ordering.
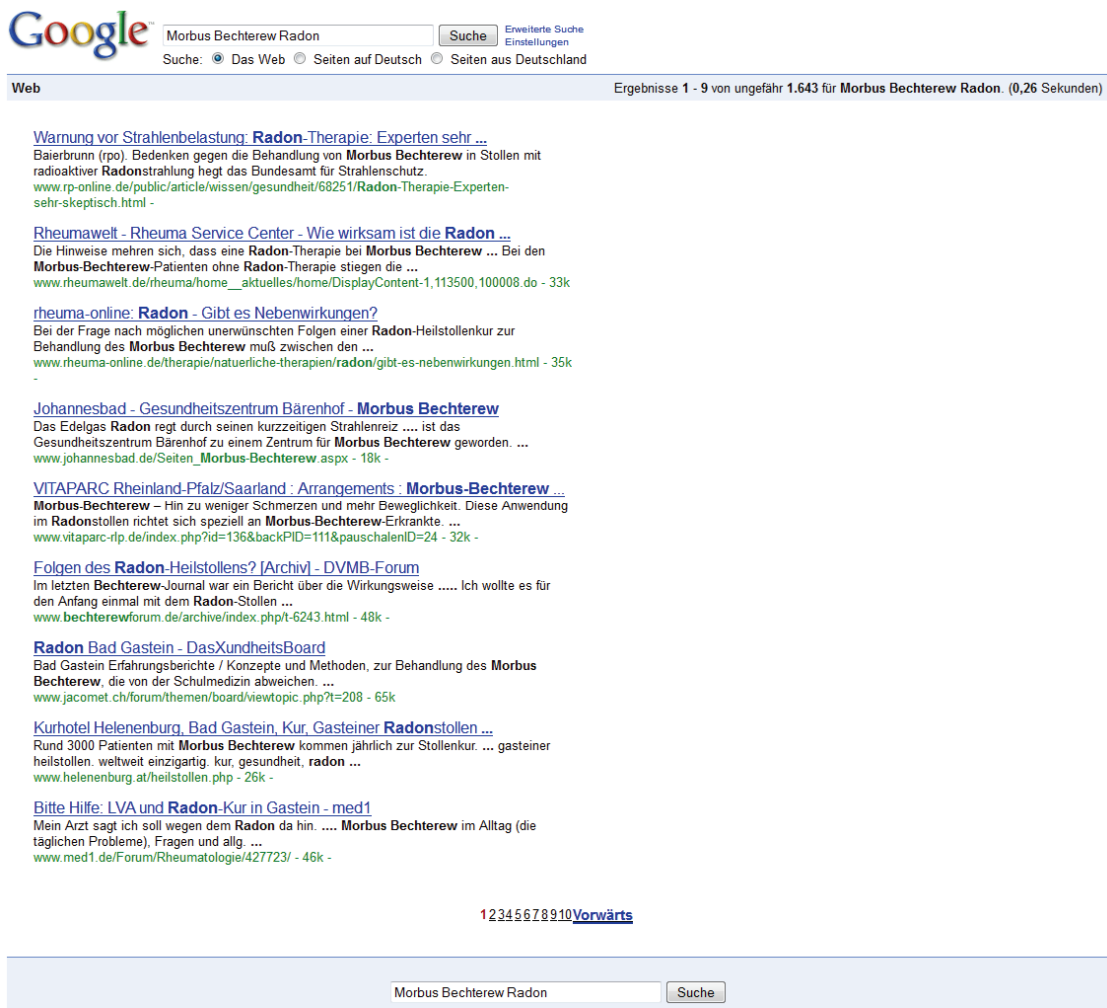
Figure 19. List interface (with search results displayed in optimal trustworthiness order).
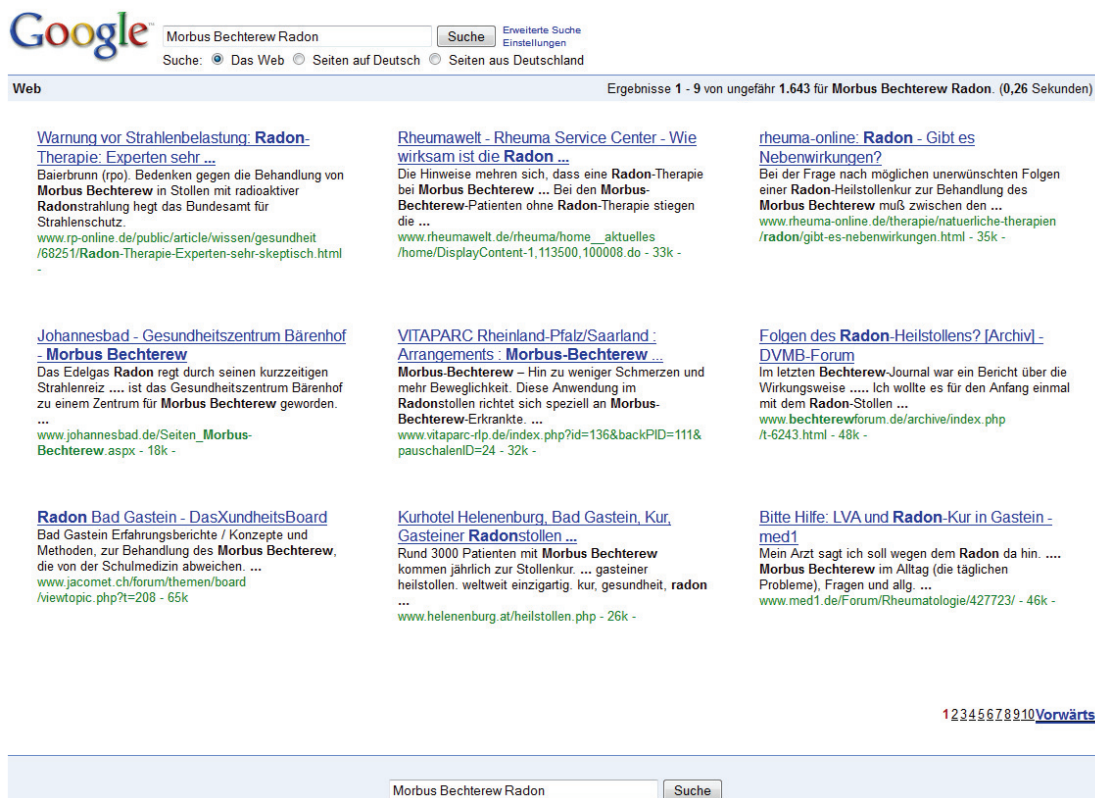
Figure 20. Grid interface (with search results displayed in optimal trustworthiness order).

As a second 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 lists with nine search results each. The search results on the lists were presented in random order to the participants. Participants' task was to order the search results of a list according to the expected trustworthiness of the corresponding Web pages from 1 = most trustworthy to 9 = least trustworthy. Based on the mean ranks of this data, two different trustworthiness orders were constructed for the SERPs: an *optimal trustworthiness order*, with the search results rated as most trustworthy being on top of the list or in the upper left corner the grid, respectively, and a *reversed trustworthiness order* with the search results rated as most trustworthy being at the bottom of the list or in the bottom right corner of the grid, respectively. Participants were randomly assigned to one of the four experimental conditions, with 20 participants serving in each of the conditions. However, as indicated above, one participant from the condition "grid interface with optimal trustworthiness order" had to be excluded from data analyses because of high prior domain knowledge.

Additionally, s*earch result trustworthiness* was considered as a third within-subjects 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 as a function of the search results trustworthiness (cf. "relevance rank" in Keane et al., 2008).

As a fourth factor participants' *epistemic beliefs* about whether or not the Web contains correct knowledge were assessed (see section 6.2.3.4 for details) and used as a continuous between-subject factor.

6.2.3 Materials and apparatus

6.2.3.1 Task

The task used in the experiment was to seek information on the WWW about two competing therapies ('radon therapy' and 'infliximab therapy') for Bechterew's disease (i.e., a chronic inflammatory rheumatic disease affecting the spine). The reason for choosing this medical topic was that it is complex and controversially discussed. Furthermore, participants' prior knowledge on this topic was expected to be low. Participants were confronted with a request from a fictitious friend (cf. Study 1), 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 (see also section 6.2.4) to conduct a Web research regarding the pros and cons of both therapies in order to provide their friend with more information about the therapies. As in Study 1 participants were informed that the therapies were controversially discussed.

6.2.3.2 Web materials

For their Web research, participants were provided with two prefabricated 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. Sample screenshots of the two SERPs are provided in Appendix D. For their Web research participants were given four minutes per SERP. All search results and 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 drug companies), and laypeople (e.g. forum pages). Furthermore, the different 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 vs. grid interface) 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, sponsored links (i.e., ads) and the hyperlinks "in cache" and "similar pages" were not included on the SERPs. Furthermore, the search results included no time data. Search result links that have already been selected were marked in purple color, whereas not yet selected links were displayed in blue. Web materials were presented with Microsoft Internet Explorer 7.

6.2.3.3 Apparatus

For eye tracking during task processing a 50 Hz Tobii 1750 remote eye tracking system with infrared-cameras built into a 17-inch monitor (www.tobii.com) was used. The Web stimulus recording mode of the ClearView 2.7.1 analysis software was used that captures not only the eye movements, but the entire task performance process (including mouse operations). Because in the present study only eye-tracking data on SERPs (i.e., mere textual information) was analyzed, the minimum fixation duration was set to 80 milliseconds (instead of 100 ms as in Study 1) with a fixation radius of 30 pixels. The viewing distance between the participants and the screen was fixed to 65 cm, using an adjustable 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, a nine-point calibration validation was applied 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 (Kammerer, 2010). Participants' retrospective verbal reports were recorded digitally

by Camtasia 3.0 software using a standard microphone attached to the PC while replaying the eye-movement recordings.

6.2.3.4 Internet-specific epistemic beliefs measure

To assess the extent to which participants believed that the Web contains correct knowledge, a translated version of the scale "certainty and source of knowledge" (8 items; Cronbach's α = .83) of the Internet-Specific Epistemological Questionnaire (ISEQ, Strømsø & Bråten, 2010) was used. As in the original questionnaire, the items were formulated with regard to Web-based knowledge in participants' field of study (i.e., study-related knowledge). The eight items had to be rated on 5-point Likert-type response scales ranging from 1 (*totally disagree*) to 5 (*totally agree*). High scores represented the belief that the Web is a reliable resource that contains correct knowledge about study-related contents. In contrast, low scores indicated that participants are more likely to doubt that the Web is a reliable resource that contains correct knowledge about study-related contents. Sample items state that the Web contains accurate knowledge in participants' field of study, the truth about almost every issue raised in participants' study courses, most of what is true in participants' field of study, or expert statements that help to resolve difficult problems in participants' field of study (for the full list of items as well as means and standard deviations of the items, see Appendix E). To avoid Web search performance being affected by thoughts provoked by the questionnaire, the questionnaire was administered online about one week after the experiment. Scores on the epistemic beliefs scale were normally distributed ($W = 0.98$, $p = .21$) with a mean of 2.36 ($SD = 0.74$) and a range from 1.13 to 3.88. There were no differences between the four experimental conditions regarding participants' epistemic beliefs (all Fs < 1) (see Table 8 for means and standard deviations).

6.2.3.5 Control variables

Demographics (gender, age), computer- and Web search experience and skills (3 items, see Appendix C; Cronbach's α = .73), and prior knowledge on Bechterew's disease or other rheumatic diseases and respective therapies (10 items, see Appendix F, Cronbach's α = .63) were assessed as control variables (see Table 8 for means and standard deviations). Except for gender and age, items had to be rated on five-point Likert-type response scales ranging from 1 (*totally disagree* or *very low*) to 5 (*totally*

agree or *very high*). ANOVAs revealed no differences between the four experimental conditions regarding gender (interface: $\chi^2(1, N = 79) = 0.58$, $p = .45$; trustworthiness order: $\chi^2(1, N = 79) = 1.72$, $p = .19$), computer- and Web search experience and skills (interface: $F < 1$; trustworthiness order: $F(1, 75) = 1.29$, $p = .26$; interface x trustworthiness order: $F(1, 75) = 2.67$, $p = .11$), and prior domain knowledge (interface: $F < 1$; trustworthiness order: $F(1, 75) = 2.07$, $p = .15$; interface x trustworthiness order: $F < 1$). However, for age marginally significant differences were revealed for interface, $F(1, 75) = 3.19$, $p = .08$ (trustworthiness order: $F < 1$; interface x trustworthiness order: $F < 1$). Therefore, age was included as covariate in all further analyses.

Table 8

Means (and Standard Deviations) of Internet-specific Epistemic Beliefs and of Control Variables as a Function of Interface and Trustworthiness Order

	Optimal trustworthiness order		Reversed trustworthiness order	
	List interface (*n* = 20)	Grid interface (*n* = 19)	List interface (*n* = 20)	Grid interface (*n* = 20)
Age	24.85 (0.82)	22.84 (0.84)	24.65 (0.82)	23.70 (0.82)
Gender	4 m, 16 f	2 m, 17 f	6 m, 14 f	5 m, 15 f
Epistemic beliefs 1 (low) – 5 (high)	2.33 (0.15)	2.25 (0.15)	2.32 (0.15)	2.27 (0.15)
Prior domain knowledge 1 (low) – 5 (high)	1.29 (0.09)	1.23 (0.09)	1.41 (0.09)	1.36 (0.09)
Computer- and Web search experience and skills 1 (low) – 5 (high)	3.52 (0.70)	3.77 (0.46)	3.58 (0.47)	3.40 (0.63)

6.2.4 Procedure

Participants were tested in individual sessions of approximately one hour. Before participants started on the search task, they were administered a computer-based questionnaire to assess the control variables. Furthermore, they received some

general instructions about the Web search experiment and were calibrated on the eye tracking system using a 9-point calibration.

Then, they underwent a training task for approximately two minutes to get acquainted with the experimental setup. This training task, which was about the weight loss methods low carb and low fat (cf. Study 1), was constructed equivalent to the subsequent main task. By means of two start Web pages with the hyperlinks "low carb" or "low fat" participants could access two preselected SERPs that according to the experimental condition were presented in the form of a list interface or a grid interface. Participants task was to inform themselves about pros and cons of the two weight loss methods.

After the training task, participants received the instructions for the main task including the fictitious request of a friend as well as a brief description of the Bechterew's disease (e.g., symptoms, causes, course of the disease) to avoid that participants spent parts of their Web search to inform themselves about these issues, rather than about the two therapies. 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 and participants had four minutes to search for information about the second therapy. Participants could access all Web pages corresponding to the 18 search results presented, but were not allowed to generate new SERPs by changing the search terms. Thus, whereas when searching on the open WWW users can repeatedly refine their search terms and have access to billions of Web pages, in the controlled experimental setting of the present study participants were provided with a restricted information space. Therefore, to ensure that participants preselected and focused on some of the information sources – which would reflect a real search situation on the WWW – rather than reading all available information, the total search time was limited to eight minutes.

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 given five minutes to write down the pros and cons of the two therapies on the basis of their Web research. Then, cued retrospective verbal protocols (cf. Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005) were obtained by presenting participants with the recordings of their eye movements (at 50% speed) and asking them each time when the screen recordings showed their eye movements on the Google SERPs to report what they were thinking during task processing. Whenever participants stopped verbalizing their thoughts, the experimenter reminded them (after approx. 5 seconds) to keep thinking aloud. This procedure was repeated for both SERPs. Finally, about one week after the experiment, participants were administered an online questionnaire to assess their epistemic beliefs.

6.2.5 Data analyses and dependent variables

6.2.5.1 Eye-tracking data

To analyze participants' eye-tracking data, for each of the nine search results on a SERP a polygonal area of interest (AOI) was defined covering the title, excerpt, and URL of a search result. It was determined for all AOIs whether, for which amount of time, and in which order a participant was looking at these areas. Although the shape of the AOIs differed between the two search interfaces (list vs. grid interface), their size, text content, font style, and font size were equal in both interfaces (see Figures 19 and 20). As no differences were expected between the two SERPs (i.e., the 'radon SERP' and the 'infliximab SERP') eye-tracking data of both SERPs were collapsed in the statistical analyses.

Viewing sequences. First of all, participants' viewing sequences on a SERP, that is, the order in which the search results (i.e., the AOIs) were visually inspected, were analyzed (including returns to a SERP after having visited a Web page). A visual inspection was defined as at least one fixation ($\geq$ 80 ms) within a search result. Specifically, the *homogeneity* and *linearity* of participants' viewing sequences in the four experimental conditions were computed as two dependent variables. 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 &

Holmes 2002; Sankoff & Kruskal, 1983). Similar strings need fewer transformations and thus have smaller distances. Participants' viewing sequences were described as strings. Each participant provided one AOI string per SERP. Revisits of an already visually inspected AOI did not count as an element of the string (cf. Joachims et al., 2005), which resulted in maximum string length of 9 per SERP. The *homogeneity* 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 19 or 18 other participants were calculated. Finally, the mean similarity percentages (i.e., the homogeneity scores) in the different conditions were compared statistically. Furthermore, the *linearity* of participants' viewing sequences on a SERP within an experimental condition was assessed by computing the similarity percentage between a participant's string and a linear string (search result 1, search result 2, search result 3, search result 4, etc.). 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, because 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 (i.e., the maximum string would be search result 1, search results 4, search result 7, search result 2, search result 5, search result 8, search result 3, search result 5, search result 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.

Visual attention to search results. The third dependent eye-tracking variable was the *total fixation time* (in milliseconds) on a search result as a function of its trustworthiness, that is, the total time for which participants during their eight-minute Web search attended to a search result of a specific trustworthiness (i.e., trustworthiness ranks 1-9). Time data was transformed into seconds for ease of interpretation.

6.2.5.2 Log file data (mouse clicks)

With regard to participants' selection data, the search results participants selected from the SERPs during their eight-minute Web search 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 variable, the *selection frequency* (i.e., 0-2 times) of a search result of a specific trustworthiness (i.e., trustworthiness ranks 1-9) was counted.

6.2.5.3 Coding scheme for thinking-aloud protocols

For the analysis of participants' cued retrospective verbalizations obtained while replaying the eye-movement recordings of the two Google SERPs, the coding scheme of Study 1 (cf. Chapter 4.2.5.1) was refined and specially-tailored to the evaluation of search results. For more fine-grained analyses of quality-related evaluation processes on SERPs, the criterion credibility of Study 1 was subdivided into three criteria differing in their level of specificity, namely, *general quality*, *type of source*, and *credibility* (from least to most specific criterion). In addition, utterances regarding the *search result position* were included in the coding scheme. Short descriptions of the resulting four evaluation criteria are provided in Table 9. Based on the findings of Study 1, up-to-dateness, structure, and intended audience were not used as criteria in the present study, as they did not seem to play a role in the evaluation of search results. Two raters familiar with the search task and the Web materials as well as with the coding scheme scored 25% of the protocols. Inter-rater reliability computed on this subsample of protocols yielded a Cohen's kappa of $k = .70$. One rater scored the remaining protocols and only the coding of this rater was used for further data analyses. Coding was realized with the software tool MEPA 4.10 (Erkens, 2005). As dependent variables the number of utterances referring to the different types of evaluation criteria was analyzed. Verbal utterances expressed across the two SERPs were aggregated in the analyses.

Table 9.

Coding Scheme for Thinking-aloud Protocols

Evaluation criteria	Short description
Search result position	The search result is selected because it is the first, the highest ranked, the topmost search result.
General quality	The search result or the corresponding web page is (or is not) good.
Type of source	The type of source, e.g., a forum, a drug company, the department of health is mentioned without an explicit credibility rating.
Credibility	The source is (or is not) expected to provide trustworthy, reliable, neutral, official information; information provided by an expert.

6.2.5.4 Quality of information problem solving

As the product of the search task, the arguments participants listed in favor and against the two therapies were analyzed. For the pool of 18 web pages, 32 different (but partially overlapping) statements which could be used as arguments for and against the two therapies were identified. Each of these arguments could be found in one or several Web page(s). Examples are "scientific studies showed significant decrease of disease activity after radon therapy", "radon stabilizes the immune system", "infliximab therapy increases the risk of infections", or "it is scientifically confirmed that infliximab reduces disease activity". Two raters familiar with the search task and the Web materials as well as with the list of arguments scored the arguments of 20 participants (25%) by classifying them as one of the 32 arguments or by identifying an argument as a false or non-argument. Inter-rater reliability computed on this subsample of protocols yielded a Cohen's kappa of $k = .81$. One rater scored the remaining argument lists and only the codings of this rater were used for further data analyses.

As a first dependent variable the *overall number of arguments recalled from the 18 Web pages* was analyzed. More important, the *number of arguments recalled from the three most trustworthy sources* was analyzed as a second dependent variable. It has to be noted, though, that these arguments could partly be found in other Web

pages as well. In addition, for exploratory analyses it was analyzed whether or not participants referred to the source of information in their arguments.

6.3   Results

An alpha level of .05 was used for the statistical tests reported. Effects were considered marginally significant when p-values were between .05 and .10. Epistemic beliefs and age were used as covariates (z-scored) in all analyses. Furthermore, interaction terms between the covariates and the experimental conditions were included in the statistical analyses[3]. As in Study 1, significant interaction effects were further examined and graphed using the procedures outlined by Aiken and West (1991). Specifically, simple comparisons were computed for low epistemic belief scores (defined as one standard deviation below the sample mean; $M - 1\ SD$) and for high epistemic belief scores (defined as one standard deviation above the sample mean; $M + 1\ SD$) as well as for younger participants ($M - 1\ SD$) and older participants ($M + 1\ SD$), using moderated regression analyses. All means and standard errors of the dependent measures reported in the following are corrected for the influence of epistemic beliefs and age.

6.3.1   Eye-tracking data

6.3.1.1   Homogeneity and linearity of viewing sequences

During the eight-minute Web search participants visually inspected nearly all ($M = 92.75\%$, $SE = 1.17$) of the 18 search results. Table 10 shows means and standard errors of the homogeneity and linearity of the viewing sequences on the SERPs as a function of interface and trustworthiness order.

---

[3] For the covariate age, only in case of heterogeneous regression slopes the interaction terms were included in the analyses.

Table 10

Means (and Standard Errors) of the Eye-tracking Data as a Function of Interface and Trustworthiness Order

Viewing sequences	Optimal trustworthiness order		Reversed trustworthiness order	
	List interface	Grid interface	List interface	Grid interface
Similarity between participants (in %)	59.90 (1.74)	39.92 (1.85)	62.93 (1.74)	37.83 (1.74)
Similarity with linear (line-by-line/top-to-bottom) string (in %)	81.15 (3.82)	53.80 (4.07)	77.80 (3.81)	46.01 (3.83)
Similarity with linear column-by-column string (in %)	NA	33.19 (3.05)	NA	32.70 (2.94)

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 ANCOVA with interface (list vs. grid) and trustworthiness order (optimal vs. reversed) as between-subject factors and epistemic beliefs and age as covariates showed a significant main effect of interface ($F(1, 70) = 150.96$, $p < .001$, partial $\eta^2 = .68$). Viewing sequences of participants using a list interface were significantly more homogeneous, with 60.91% similarity ($SE = 1.23$), than were those of participants using a grid interface, with 38.87% similarity ($SE = 1.28$). Besides this, there were neither main effects of trustworthiness order ($F < 1$), epistemic beliefs ($F < 1$), or age ($F(1, 70) = 2.25$, $p = .14$), nor were there significant interactions between interface and trustworthiness order ($F(1, 70) = 1.37$, $p = .25$), interface and epistemic beliefs ($F < 1$), trustworthiness order and epistemic beliefs ($F < 1$), or between all three factors ($F(1, 70) = 1.26$, $p = .27$).

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 ANCOVA also showed a significant main effect of interface ($F(1, 70) = 56.36$, $p < .001$, partial $\eta^2 = .45$). Viewing sequences of participants using a list interface had a similarity of 79.48% ($SE = 2.71$)

with the linear string, whereas those of participants using the grid interface only had a similarity of 49.91% (SE = 2.81) with the linear string. Besides this, there were neither main effects of trustworthiness order ($F(1, 70)$ = 2.08, p = .15), epistemic beliefs, or age, nor were there any significant two- or three-way interactions between the factors (all Fs < 1). Additionally, for the grid interface conditions the similarity percentage between participants' string and a linear column-by-column string was calculated. An ANCOVA for the data of the 39 grid interface users with trustworthiness order as between-subject factor, the type of string (line-by-line vs. column-by-column) as within-subject factor and epistemic beliefs and age as covariates showed that participants' viewing sequences had an even lower similarity of 32.85% (SE = 2.86) with the column-by-column string than with the line-by-line string ($F(1, 34)$ = 15.58, p < .001, partial η^2 = .32). There were no main effects of trustworthiness order ($F(1, 34)$ = 1.74, p = .20), epistemic beliefs, or age, and no significant interactions between these factors (all Fs < 1).

Furthermore, to analyze the heterogeneity of grid interface users' viewing sequences in greater detail, a hierarchical cluster analysis using the Ward method with the two grouping variables 'line-by-line similarity percentage' and 'column-by-column similarity percentage' was conducted. The cluster analysis identified two subgroups of grid interface users: The first group that comprised the majority (n = 34) of the participants (16 in the optimal and 18 in the reversed trustworthiness order), showed a rather low similarity percentage with both the line-by-line string (M = 44.18%) and the column-by-column string (M = 34.25%). The second group, comprising only five participants (three in the optimal and two in the reversed trustworthiness order), showed a high similarity percentage with the line-by-line string (M = 80.69%) and a low similarity percentage with the column-by-column string (M = 27.20%).

6.3.1.2 Total fixation time on search results as a function of search result trustworthiness

Means and standard errors of the total fixation time on search results as a function of search result trustworthiness and interface are depicted in Figure 21 for the optimal trustworthiness order and in Figure 22 for the reversed trustworthiness order.
An ANCOVA with interface (list vs. grid) and trustworthiness order (optimal vs. reversed) as between-subject factors, search result trustworthiness (trustworthiness rank 1-9) as within-subject factor, and epistemic beliefs and age as covariates

showed no significant main effects of interface (F(1, 70) = 2.45, p = .12), trustworthiness order (F(1, 70) = 1.00, p = .32), or search result trustworthiness (F(4.80, 336.17) = 1.66, p = .15; Greenhouse-Geisser corrected) on the total fixation time on search results. However, there was a significant main effect of epistemic beliefs (F(1, 70) = 4.04, p = .05, partial η^2 = .06). The higher participants' beliefs that the Web contains correct knowledge the shorter they fixated on search results (β = -.23, p = .05). There was no significant main effect of age (F < 1). Furthermore, there were significant two-way interactions between interface and trustworthiness order (F(1, 70) = 3.99, p = .05, partial η^2 = .05) and between trustworthiness order and search result trustworthiness (4.80, 336.17) = 17.90, p < .001, partial η^2 = .20; Greenhouse-Geisser corrected). However, these interactions have to be interpreted in the light of a significant three-way interaction between all three factors (F(4.80, 336.17) = 2.15, p = .05, partial η^2 = .03; Greenhouse-Geisser corrected) that will be explained below. There were no significant interactions between interface and search result trustworthiness (F(4.80, 350.46) = 1.47, p = .20; Greenhouse-Geisser corrected), or between epistemic beliefs and the experimental conditions (all Fs < 1.42).

For the optimal trustworthiness order (see Figure 21) Bonferroni posthoc tests revealed no significant differences between the two interfaces with respect to participants' total fixation time on search results of any of the nine trustworthiness ranks. In contrast, in the reversed trustworthiness order (see Figure 22) participants using the list interface attended significantly longer to the five least trustworthy search results than participants using the grid interface. In other words, with a list interface the top search results (i.e., in the optimal trustworthiness order the most trustworthy ones and in the reversed trustworthiness order the least trustworthy ones) received significantly more attention than the lower ones. In contrast, with a grid interface in both the optimal and reversed trustworthiness order, total fixation time on 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 optimal trustworthiness order, and rank 9 marginally longer than rank 8 in the reversed trustworthiness order). That is, with a grid interface nearly all search results on a SERP were attended to equivalently long.
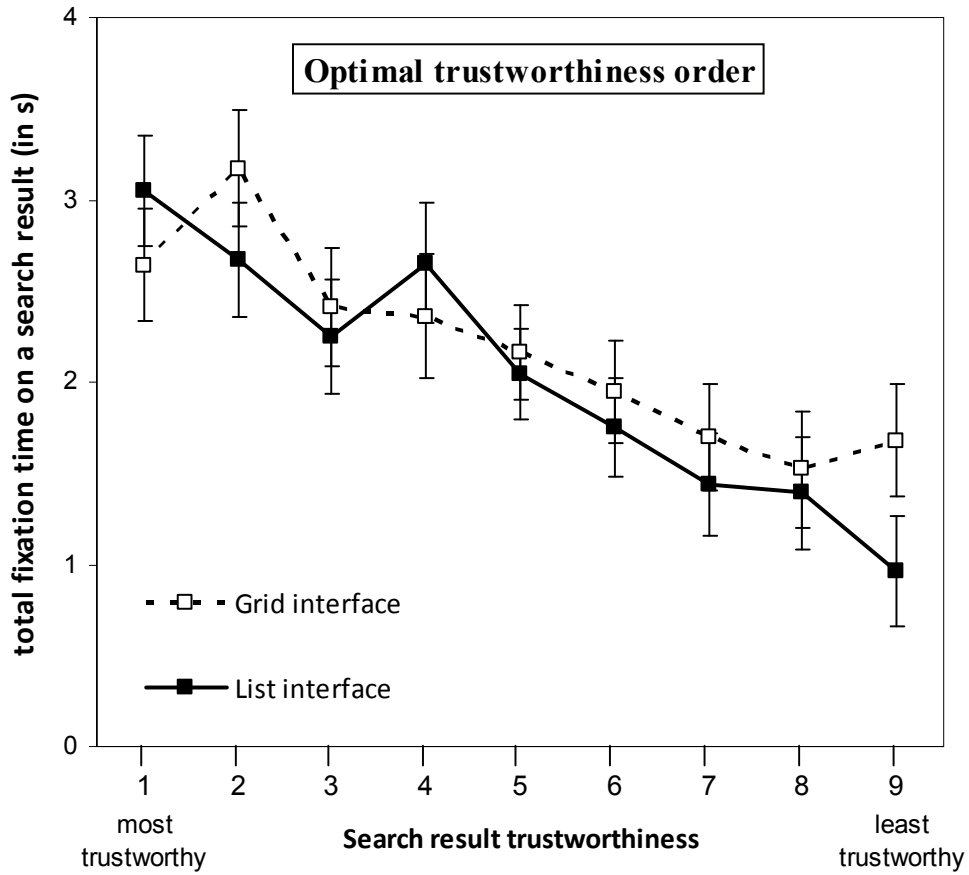
Figure 21. Total fixation time (means and standard errors) on search results as a function of search result trustworthiness and interface for the optimal trustworthiness order.
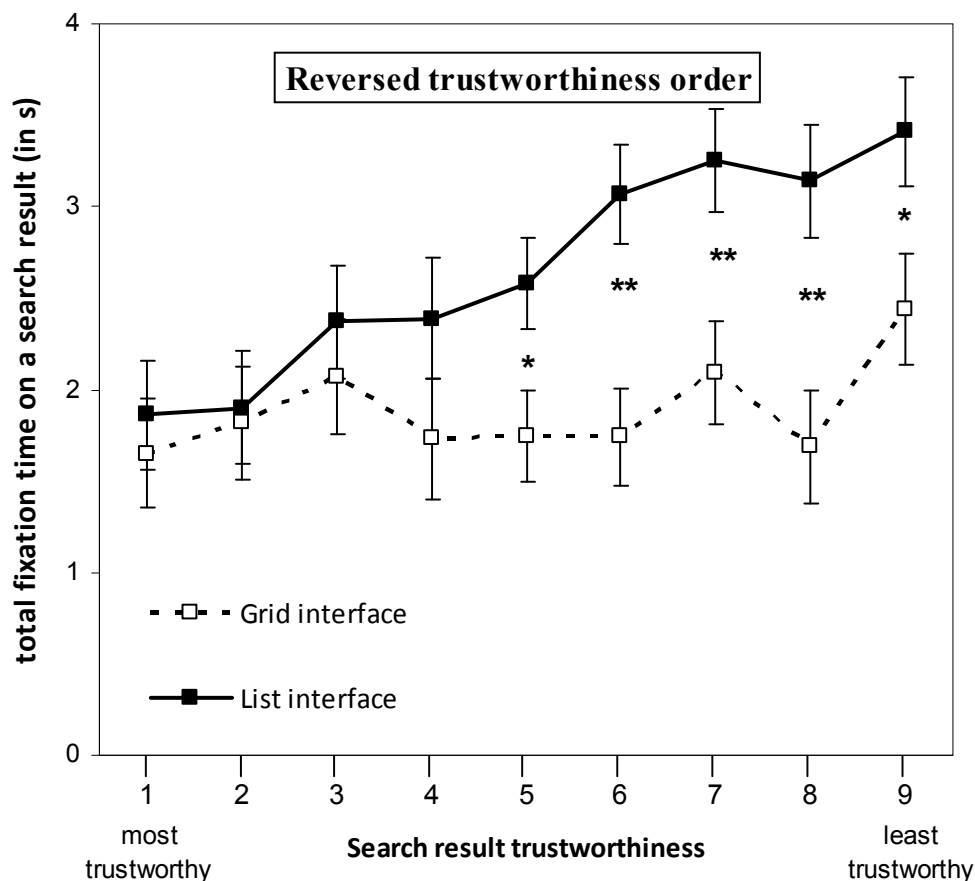
Figure 22. Total fixation time (means and standard errors) on search results as a function of search result trustworthiness and interface for the reversed trustworthiness order. ** $p \leq .01$, * $p \leq .05$

6.3.2 Search result selection

During the eight-minute Web search participants selected 51.38% ($SE = 1.53$) of the 18 search results on average. Means and standard errors of the selection frequency of search results as a function of search result trustworthiness and interface are depicted in Figure 23 for the optimal trustworthiness order and in Figure 24 for the reversed trustworthiness order.

An ANCOVA with interface (list vs. grid) and trustworthiness order (optimal vs. reversed) as between-subject factors, search result trustworthiness (trustworthiness rank 1-9) as within-subject factor, and epistemic beliefs and age as covariates showed no significant main effects of interface ($F(1, 70) = 2.44$, $p = .12$), trustworthiness order ($F(1, 70) = 1.50$, $p = .23$), epistemic beliefs ($F(1, 73) = 1.00$, $p = .32$), or age ($F(1, 70) = 2.70$, $p = .11$) on the selection frequency of search results.

There was a significant main effect of search result trustworthiness ($F(6.79, 475.43)$ = 28.38, $p < .001$, partial $\eta^2 = .29$) and a significant two-way interaction between trustworthiness order and search result trustworthiness ($F(6.79, 475.43) = 4.03$, $p < .001$, partial $\eta^2 = .05$). However, these effects have again to be interpreted in the light of a significant three-way interaction between interface, trustworthiness order, and search result trustworthiness ($F(6.79, 475.43) = 2.08$, $p = .05$, partial $\eta^2 = .03$) that will be explained below. There were no significant interactions between interface and search result trustworthiness ($F(6.79, 475.43) = 1.07$, $p = .38$; Greenhouse-Geisser corrected), between interface and trustworthiness order ($F < 1$), or between epistemic beliefs and the experimental conditions (all Fs < 1.34).

Bonferroni posthoc tests revealed that – similar to the pattern for the total fixation time - in the optimal trustworthiness order (see Figure 23) 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). In contrast, in the reversed trustworthiness order (see Figure 24) participants using the grid interface selected the three most trustworthy search results significantly more often than participants using the list interface. In other words, whereas with a list interface the three most trustworthy search results were selected significantly less often in the reversed trustworthiness order than in the optimal trustworthiness order, with a grid interface there were no differences in search results selection with respect to trustworthiness order.

Finally, it should be noted that the selection frequency of the fifth trustworthiness rank (which is the same search result in both trustworthiness orders) in all experimental conditions was remarkably low. This might be due to the fact that in both SERPs the search result of the fifth trustworthiness rank linked to a commercial Web page.
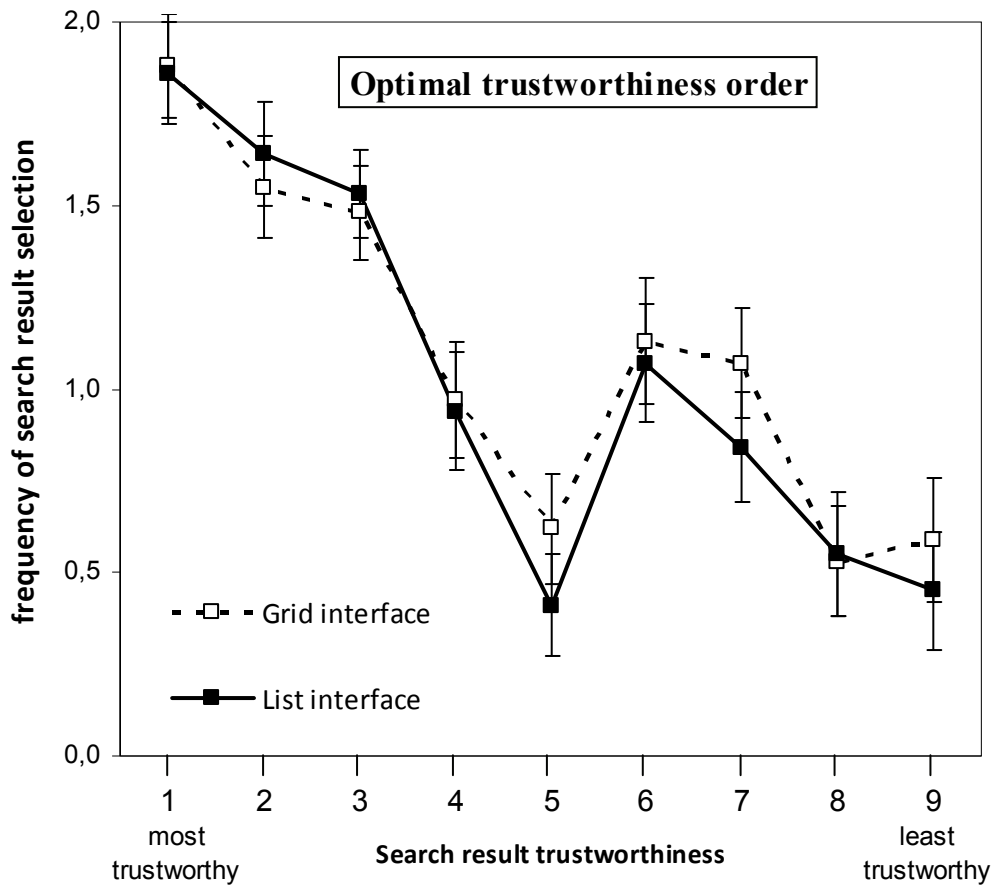
Figure 23. Frequency of search result selection (means and standard errors) as a function of search result trustworthiness and interface for the optimal trustworthiness order.
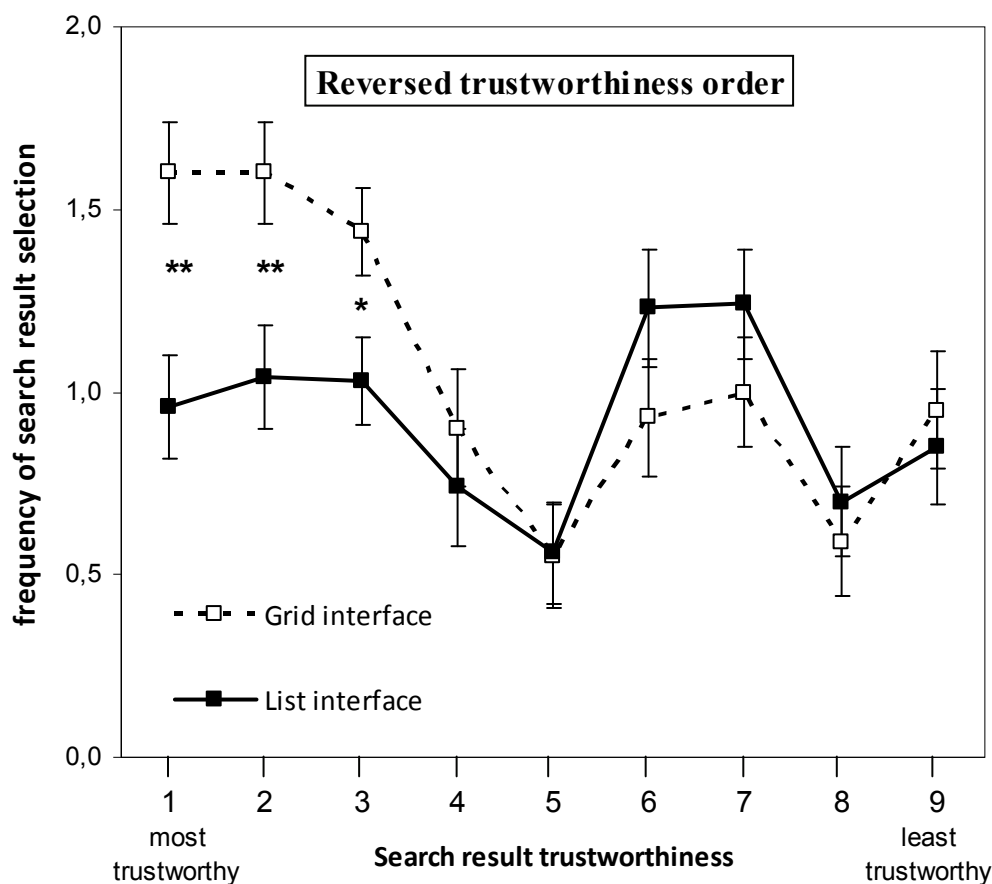
Figure 24. Frequency of search result selection (means and standard errors) as a function of search result trustworthiness and interface for the reversed trustworthiness order. ** $p \leq .01$, * $p \leq .05$

6.3.3 Number of verbal utterances

Table 11 shows means and standard errors of the number of verbal utterances related to the four evaluation criteria as a function of interface and trustworthiness order.

ANCOVAs with interface (list vs. grid) and trustworthiness order (optimal vs. reversed) as between-subject factors and epistemic beliefs and age as covariates were conducted for the number of utterances related to the four evaluation criteria *search result position*, *general quality*, *type of source*, and *credibility*.

Table 11

Means (and Standard Errors) of the Number of Verbal Utterances related to the Four Evaluation Criteria as a Function of Interface and Trustworthiness Order

Evaluation criteria	Optimal trustworthiness order		Reversed trustworthiness order	
	List interface	Grid interface	List interface	Grid interface
Search result position	2.18 (0.32)	1.21 (0.34)	0.59 (0.32)	0.18 (0.32)
General quality	0. 40 (0.16)	1.04 (0.17)	0.00 (0.16)	0.31 (0.16)
Type of source	4.12 (0.61)	4.31 (0.67)	4.41 (0.61)	5.68 (0.62)
Credibility	0.97 (0.35)	1.53 (0.38)	1.28 (0.35)	1.55 (0.35)

6.3.3.1 Number of utterances related to the search result position

With regard to the number of utterances related to the *search result position*, the ANCOVA showed significant main effects of interface ($F(1, 69) = 4.45$, $p = .04$, partial $\eta^2 = .06$) and trustworthiness order ($F(1, 69) = 16.61$, $p < .001$, partial $\eta^2 = .19$), but no main effects of epistemic beliefs and age (both Fs < 1). However, the main effects of interface and trustworthiness order on the number of utterances related to the *search result position* have to be interpreted in the light of significant or marginally significant two-way interactions with the covariates; namely a significant interaction between trustworthiness order and age ($F(1, 69) = 3.85$, $p = .05$, partial $\eta^2 = .05$) and a marginally significant interaction between interface and epistemic beliefs ($F(1, 69) = 2.79$, $p = .10$, partial $\eta^2 = .04$).

In order to probe the interaction between trustworthiness order and age, simple comparisons of trustworthiness order (optimal vs. reversed) on different levels of age were conducted following the procedure outlined by Aiken and West (1991). Results showed that younger participants (M - 1 SD) verbally addressed the search result position significantly more often in the optimal trustworthiness order than in the reversed trustworthiness order ($\Delta M = 1.86$, $SE = 0.42$, $t(69) = 1.81$, $\beta = -.59$, $p < .001$). Older participants ($M + 1$ SD), on the contrary, did not differ in the number of utterances related to the search result position in the two trustworthiness order conditions ($\Delta M = -0.62$, $SE = 0.42$, $t(69) = 1.81$, $\beta = -.20$, $p = .19$). Figure 25 represents this interaction graphically.
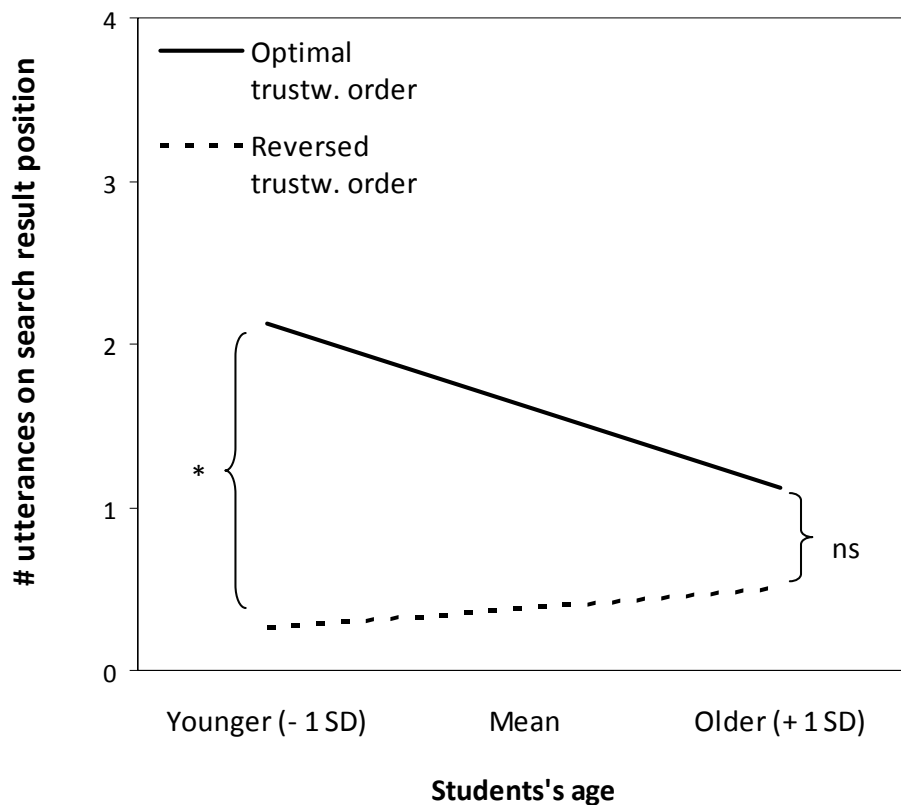
Figure 25. Interaction of trustworthiness order and age with regard to the number of verbal utterances related to the search result position. * $p \le .05$

With regard to the interaction between interface and epistemic beliefs, simple comparison revealed that students with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) verbally addressed the search result position significantly less in the grid interface than in the list interface ($\Delta M = -1.24$, $SE = 0.46$, $t(69) = -2.71$, $\beta = -.39$, $p = .01$). Students with low beliefs that the Web contains correct knowledge ($M - 1\ SD$), on the contrary, did not differ in the number of utterances related to the type of source in the two interface conditions ($\Delta M = -0.14$, $SE = 0.45$, $t(69) = -0.31$, $\beta = -.05$, $p = .76$). Figure 26 represents this interaction graphically.
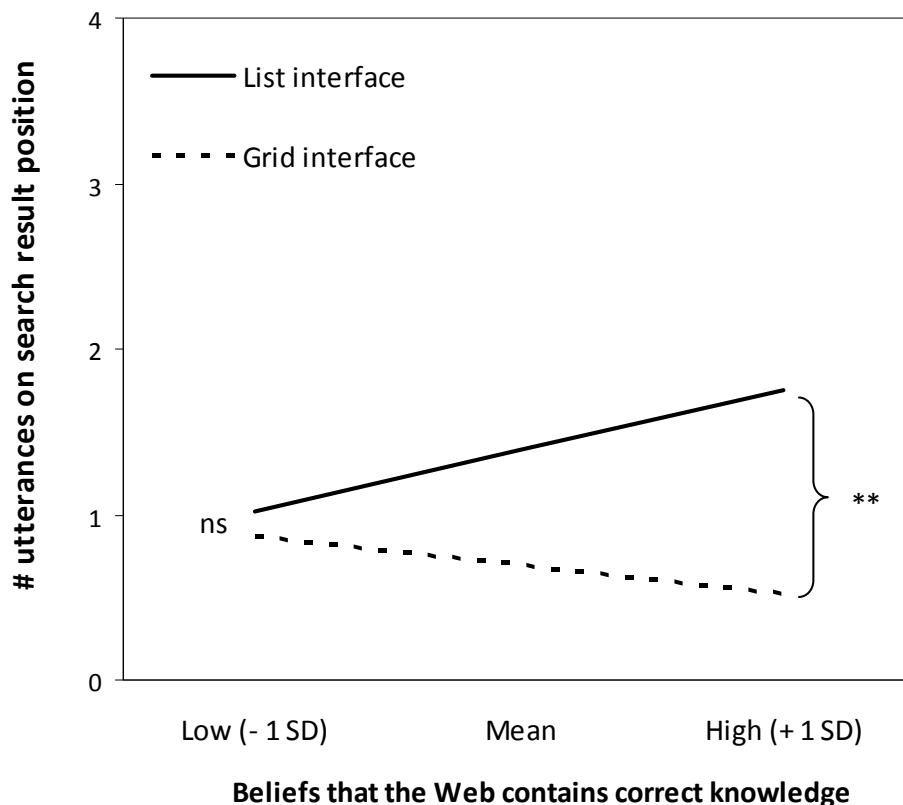
Figure 26. Interaction of trustworthiness order and epistemic beliefs with regard to the number of verbal utterances related to the search result position. ** $p \leq .01$

Besides, neither the interaction between interface and trustworthiness order, nor between trustworthiness order and epistemic beliefs, or the interaction between all three factors was significant (all Fs < 1).

6.3.3.2 Number of utterances related to the general quality

With regard to the number of utterances related to the *general quality*, the ANCOVA showed significant main effects of interface $F(1, 70) = 8.40$, $p < .01$, partial $\eta^2 = .11$) and trustworthiness order $F(1, 70) = 12.42$, $p = .001$, partial $\eta^2 = .15$), but no main effects of epistemic beliefs ($F(1, 70) = 2.10$, $p = .15$) and age ($F < 1$). With regard to the main effect of interface, participants using the grid interface showed significantly more utterances ($M = 0.68$, $SE = 0.12$) than participants using the list interface ($M = 0.20$, $SE = 0.11$). The main effect of trustworthiness order has to be interpreted in the light of a marginally significant interaction between trustworthiness order and epistemic beliefs ($F(1, 70) = 3.32$, $p = .07$, partial $\eta^2 = .05$).

Simple comparisons revealed that students with low beliefs that the Web contains correct knowledge (M - 1 SD) verbally addressed the general quality significantly less in the reversed trustworthiness order than in the optimal trustworthiness order (ΔM = -0.86, SE = 0.23, $t(70)$ = -3.73, β = -.56, $p < .001$). Students with high beliefs that the Web contains correct knowledge (M + 1 SD), on the contrary, did not differ in the number of utterances related to the general quality in the two trustworthiness order conditions (ΔM = -0.27, SE = 0.23, $t(70)$ = -1.18, β = -.17, $p = .24$). Figure 27 represents this interaction graphically.
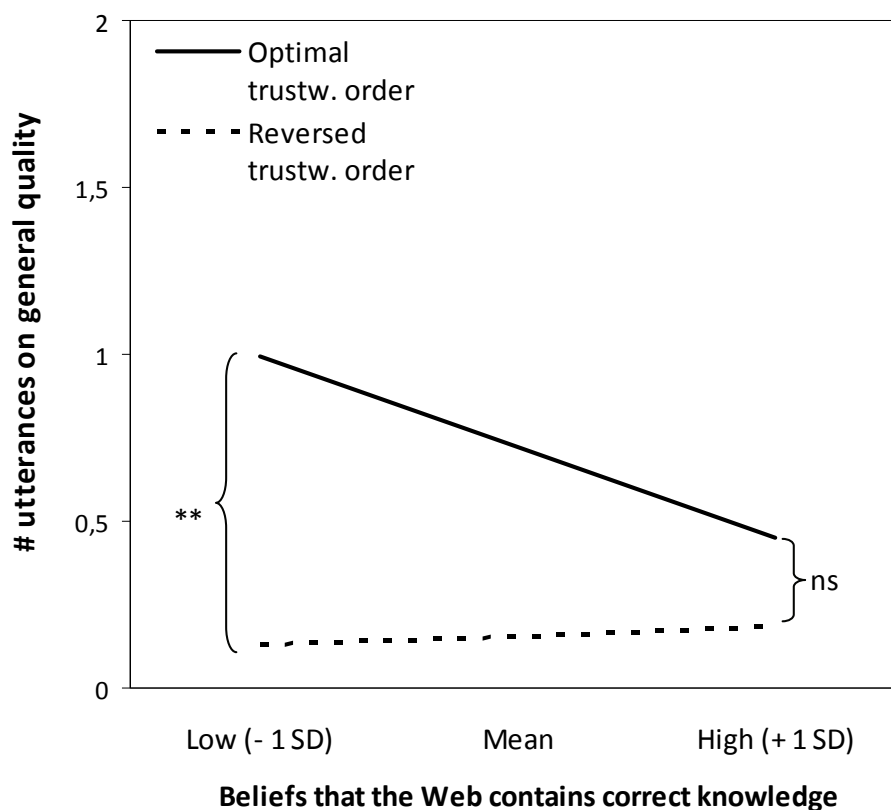

Figure 27. Interaction of trustworthiness order and epistemic beliefs with regard to the number of verbal utterances related to the general quality. ** $p \leq .01$

Besides, there were no significant interactions between interface and trustworthiness order ($F(1, 70)$ = 1.12, $p = .29$), interface and epistemic beliefs ($F < 1$), or between all three factors ($F (1, 70)$ = 1.13, $p = .29$).

6.3.3.3 Number of utterances related to the type of source

With regard to the number of utterances related to the *type of source*, the ANCOVA showed no significant main effects of interface ($F(1, 69) = 1.32, p = .25$) or trustworthiness order ($F(1, 69) = 1.80, p = .18$).

However, the main effect of epistemic beliefs was significant ($F(1, 69) = 3.87, p = .05$, partial $\eta^2 = .05$) as well as the interaction between epistemic beliefs and trustworthiness order ($F(1, 69) = 5.37, p = .02$, partial $\eta^2 = .07$). In order to probe this interaction, simple comparisons of trustworthiness order (optimal vs. reversed) on different levels of epistemic beliefs were conducted. Results revealed that students with low beliefs that the Web contains correct knowledge ($M - 1\,SD$) verbally addressed the type of source significantly more in the reversed trustworthiness order than in the optimal trustworthiness order ($\Delta M = 2.21, SE = 0.88, t(69) = 2.50, \beta = .36, p = .02$). Students with high beliefs that the Web contains correct knowledge ($M + 1\,SD$), on the contrary, did not differ in the number of utterances related to the type of source in the two trustworthiness order conditions ($\Delta M = -0.62, SE = 0.86, t(69) = 1.81, \beta = -.10, p = .47$). Figure 28 represents this interaction graphically.
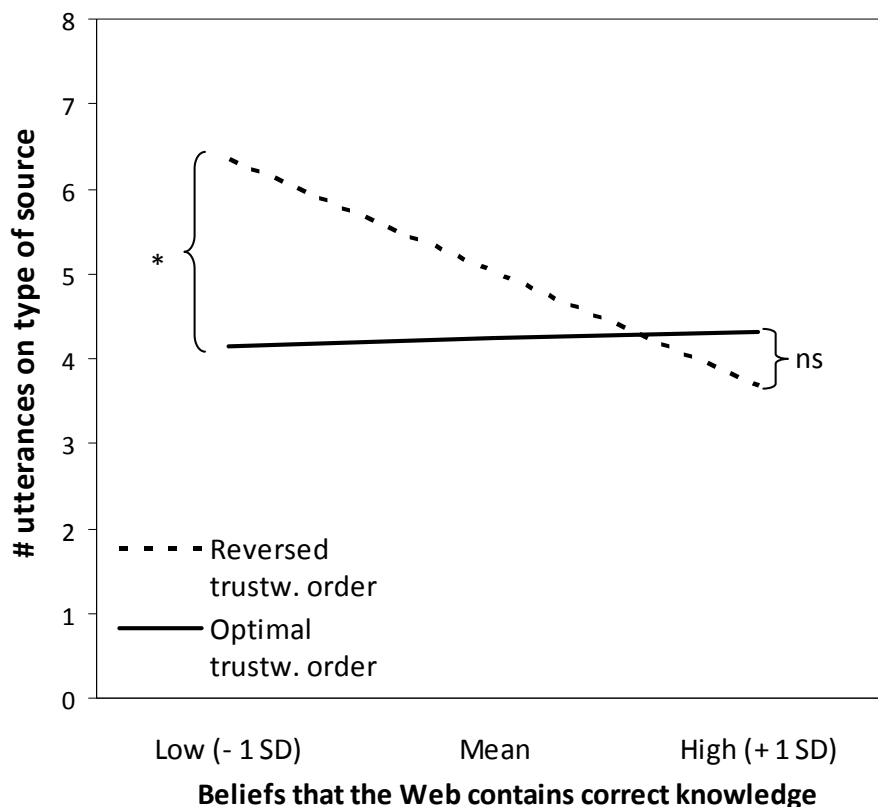
Figure 28. Interaction of trustworthiness order and epistemic beliefs with regard to the number of verbal utterances related to the type of source. * $p \leq .05$

Furthermore, there was a significant main effect of age ($F(1, 69) = 5.16$, $p = .03$, partial $\eta^2 = .07$) and a significant two-way interaction between interface and age ($F(1, 69) = 10.33$, $p < .01$, partial $\eta^2 = .13$).

Simple comparisons revealed that older participants ($M + 1$ SD) verbally addressed the type of source significantly more in the grid interface than in the list interface ($\Delta M = 2.82$, $SE = 0.96$, $t(69) = 1.81$, $\beta = .47$, $p < .01$). Younger participants ($M - 1$ SD), on the contrary, did not differ in the number of utterances related to the type of source in the two interface conditions ($\Delta M = -1.40$, $SE = 0.88$, $t(69) = 1.81$, $\beta = -.23$, $p = .12$). Figure 29 represents this interaction graphically.
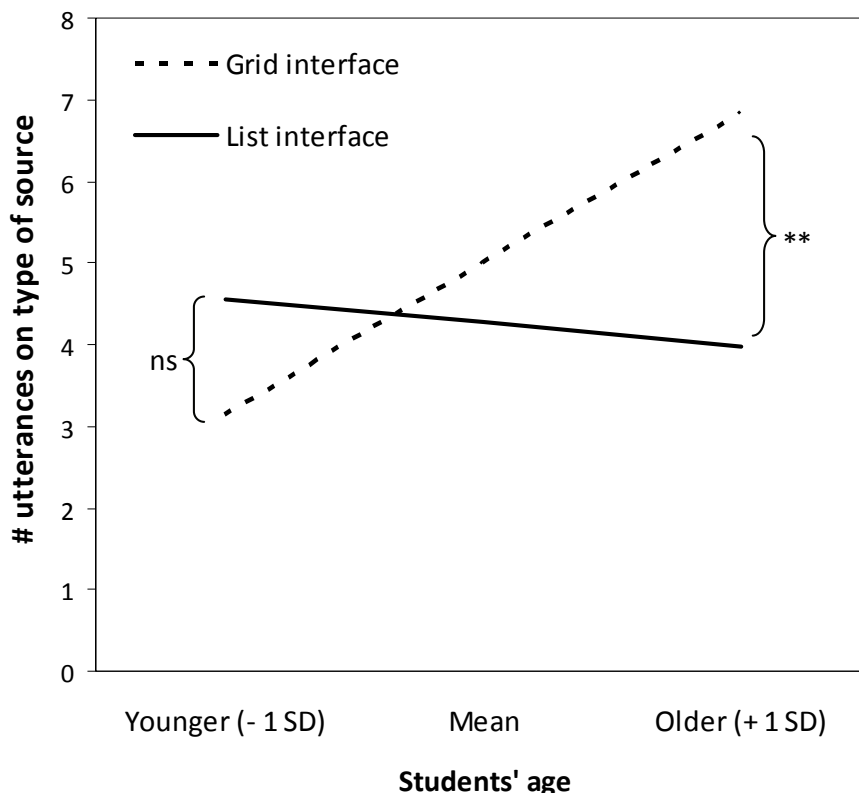
Figure 29. Interaction of interface and age with regard to the number of verbal utterances related to the type of source. ** $p \leq .01$

Besides, there were no significant interactions between interface and trustworthiness order, interface and epistemic beliefs, or between all three factors (all Fs < 1).

6.3.3.4 Number of utterances related to credibility

Finally, with regard to the number of utterances related to the *credibility* of information sources, the ANCOVA showed no main effects of interface ($F(1, 70) =$ 1.30, $p = .26$), trustworthiness order ($F < 1$), or epistemic beliefs ($F < 1$).

However, there was a significant main effect of age ($F(1, 70) = 4.30$, $p = .04$, partial $\eta^2 = .06$). The older participants, the more they addressed the credibility of information sources when evaluating search results ($\beta = .25$, $p = .04$). Besides, there were no interaction effects between interface and trustworthiness order, interface and epistemic beliefs, trustworthiness and epistemic beliefs, or between all three factors (all Fs < 1).

6.3.4 Quality of information problem solving

Table 12 shows means and standard errors of the search outcome variables as a function of interface and trustworthiness order.

Table 12

Means (and Standard Errors) of the Search Outcome Variables as a Function of Interface and Trustworthiness Order

Search outcome	Optimal trustworthiness order		Reversed trustworthiness order	
	List interface	Grid interface	List interface	Grid interface
# arguments from all Web pages	7.25 (0.52)	7.76 (0.54)	5.65 (0.52)	6.72 (0.52)
# arguments from the three most trustworthy sources	5.62 (0.45)	5.20 (0.48)	3.67 (0.45)	5.06 (0.45)

With regard to the overall number of arguments recalled from the Web pages, an ANCOVA with interface (list vs. grid) and trustworthiness order (optimal vs. reversed) as between-subject factors and epistemic beliefs and age as covariates showed a significant main effect of trustworthiness order ($F(1, 70) = 4.74$, $p = .03$, partial $\eta^2 = .06$), with participants in the optimal trustworthiness order having listed significantly more correct arguments ($M = 7.61$, $SE = 0.38$) than participants in the reversed trustworthiness order ($M = 6.44$, $SE = 0.37$). There were neither main effects of interface, epistemic beliefs, or age (all Fs < 1), nor any significant interactions between interface and trustworthiness order ($F < 1$), or between epistemic beliefs and experimental conditions (interface x epistemic beliefs: $F(1, 70) = 2.18$, $p = .15$; trustworthiness order x epistemic beliefs: $F < 1$; three-way interaction: $F < 1$) .

More important, with regard to the number of arguments listed from the three most trustworthy sources, the ANCOVA showed a significant main effect of trustworthiness order ($F(1, 70) = 5.27$, $p = .03$, partial $\eta^2 = .07$), with participants in the optimal trustworthiness order having listed significantly more of these arguments ($M = 5.41$, $SE = 0.33$) than participants in the reversed trustworthiness order ($M = 4.36$, $SE = 0.32$). There were no significant main effects of interface ($F(1, 70) = 1.10$,

$p = .30$), epistemic beliefs ($F < 1$), or age ($F < 1$). However, the ANCOVA showed a significant interaction effect between interface and trustworthiness order ($F(1, 70) = 4.00$, $p = .05$, partial $\eta^2 = .05$).

Bonferroni posthoc tests revealed that - in line with the selection data - in the optimal trustworthiness order there were no significant differences between the two interfaces with respect to the number of arguments of the three most trustworthy sources ($p = .52$). In contrast, in the reversed trustworthiness order participants using the grid interface listed significantly more of these arguments than participants using the list interface ($p = .03$). In other words, whereas with a list interface arguments from the three most trustworthy sources were listed significantly less often in the reversed trustworthiness order than in the optimal trustworthiness order ($p < .01$), with a grid interface the number of arguments listed from the three most trustworthy sources did not differ with respect to trustworthiness order ($p = .84$).

Furthermore, the ANCOVA showed a marginally significant three-way interaction between trustworthiness order, interface, and epistemic beliefs on the number of arguments recalled from the three most trustworthy sources ($F(1, 70) = 3.52$, $p = .07$, partial $\eta^2 = .05$). The two-way interactions between epistemic beliefs and interface or between epistemic beliefs and trustworthiness order were not significant ($F(1, 70) = 1.42$, $p = .24$, and F < 1, respectively). To probe the three-way interaction, simple comparisons of interface (list vs. grid) and trustworthiness order (optimal vs. reversed) and their interaction (interface x trustworthiness order) were conducted on different levels of epistemic beliefs. Results revealed that only for participants with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) the interaction between interface and trustworthiness order on the number of arguments recalled from the three most trustworthy sources was significant ($\beta = .43$, $p = .01$), but not for participants with low beliefs that the Web contains correct knowledge ($M - 1\ SD$), $\beta = .01$, $p = .95$. Specifically, in the reversed trustworthiness order, participants with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) who used the grid interface included significantly more of these arguments than those who used the list interface ($\Delta M = 2.84$, $SE = 0.90$, $t(35) = 1.81$, $\beta = .69$, $p < .01$). In contrast, participants with low beliefs that the Web contains correct knowledge ($M - 1\ SD$) did not differ between the two interfaces in the number of arguments they listed from the three most trustworthy sources ($\Delta M = 0.00$, $SE = 0.82$, $t(35) = 1.81$, $\beta = .00$, $p = .99$). Figure 30 represents this interaction graphically.
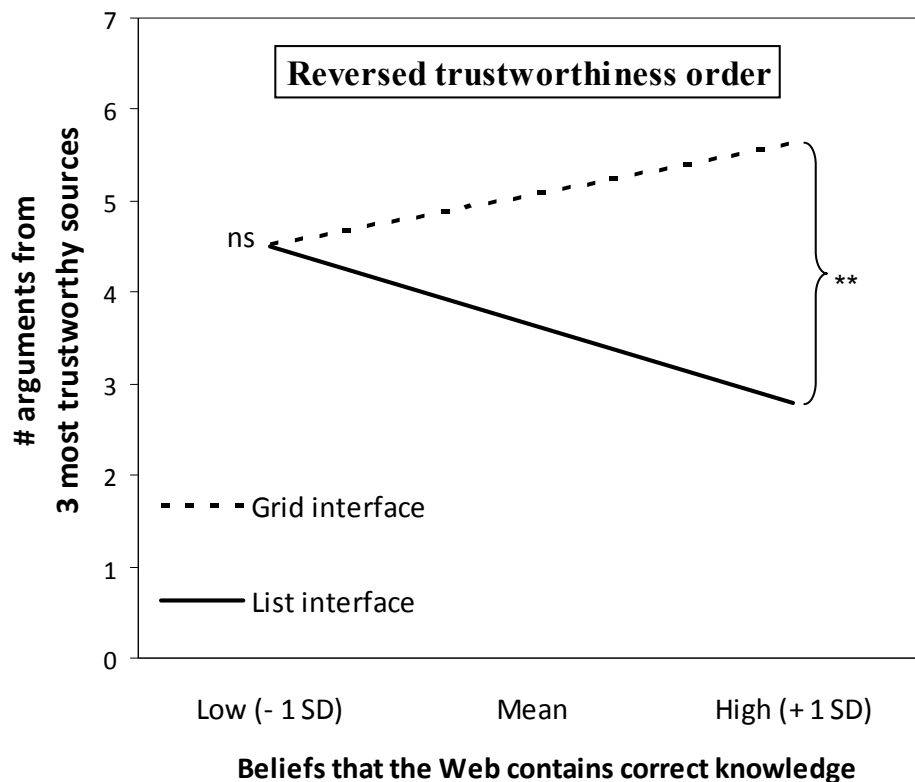
Figure 30. Interaction of interface and epistemic beliefs with regard to the number of arguments from the three most trustworthy sources for the reversed trustworthiness order. ** $p \leq .01$

In the optimal trustworthiness order, on the other hand, neither participants with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) nor those with low beliefs about this issue (M - 1 SD) differed between the two interfaces in the number of arguments they listed from the three most trustworthy sources (ΔM = -0.76, SE = 0.94, $t(34)$ = 1.81, β = -.19, p = .42, and ΔM = -0.14, SE = 1.06, $t(34)$ = 1.81, β = -.03, p = .90, respectively).

Finally, exploratory analyses showed no differences between search interfaces ($\chi^2(1, N = 79)$ = 0.84, p = .36) or trustworthiness orders with regard to whether or not participants referred to the source of information in their arguments ($\chi^2(1, N = 79)$ = 0.65, p = .42). In the optimal trustworthiness order two list interface users (10%) and four grid interface users (21.05%) referred to source information in their arguments. Similarly, in the reversed trustworthiness order four list interface users (20%) and five grid interface users (25%) referred to source information in their arguments.

6.4 Discussion

The purpose of the present study was to examine the role of both the search interface and Internet-specific epistemic beliefs in facilitating or hampering low-knowledge searchers' evaluations of information quality during Web search on a complex medical issue. Specifically, it was assumed that a search interface that refrains from making the ranking of search results the most salient feature would facilitate the evaluation of information quality on SERPs as compared to a list format that imposes a high salience on the ranking of the search results. Therefore, the present study compared a standard Google-like list interface to a grid interface with search results presented in a three-by-three grid. Moreover, a combined effect of the grid interface and doubts with regard to the quality of information on the Web was assumed, with the advantage of the grid interface over the list interface being larger for searchers with a high level of doubts that the Web contains correct knowledge about study-related contents than for searchers with few doubts (i.e., high beliefs) about this issue. The extent to which participants believed that the Web contains correct knowledge about study-related contents were assessed by the "certainty and source of knowledge" scale of the ISEQ (cf. Strømsø & Bråten, 2010).

As expected, differences between search interfaces were found for eye-tracking data (Hypotheses 1 and 2b), search result selection (Hypothesis 3a), and verbal utterances (Hypothesis 4a), but not for participants' search outcome (Hypothesis 5a). Furthermore, however, the assumption that the grid interface would particularly stimulate searchers with a high epistemic awareness to engage in source evaluations was not confirmed: Neither for eye-tracking data (Hypothesis 2c), nor for search result selection (Hypothesis 3b), or verbal data related to source information (Hypothesis 4b) interaction effects between interface and epistemic beliefs were found in the present study. For verbal utterances related to the search result position (Hypothesis 4c) and for participants' search outcome (Hypothesis 5b) interaction effects were found, however, in the reversed direction, namely that the grid interface compensated for a low level of doubts (i.e., high beliefs) with regard to quality of Web information. Finally, relationships of epistemic beliefs with eye-tracking data (Hypothesis 2a), as well with verbal utterances were found. The following sections will discuss in detail the results of this study with regard to the effects of the search

interface (6.4.1) and of epistemic beliefs (6.4.2) on participants' source evaluations on SERPs.

6.4.1 Effects of the search interface on the evaluation of information quality

First of all, in line with Hypothesis 1, the effects of the salience of the ranking of the search interface were shown in participants' viewing sequences. Grid interface users attended to the search results presented on a SERP in a less linear (i.e., line-by-line or column-by-column) sequence than list interface users. Furthermore, viewing sequences among participants were more heterogeneous in the grid interface than in the list interface. 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 interface supports a more free exploration of the search results than a list interface.

Second, the findings with regard to the search interface indicate that, in line with Hypothesis 2b and 3a, a grid interface stimulates searchers to base their selection decisions less on the position and more on own evaluations of the search results. When the top search results were of low trustworthiness (reversed trustworthiness order) students using the grid interface fixated significantly shorter on the five least trustworthy search results and selected the three most trustworthy search results significantly more often than students using the list interface. In contrast, in the optimal trustworthiness order no differences were found between the two interfaces regarding users' eye-tracking and selection data.

It has to be noted, however, that, contrary to Hypothesis 2b and 3a, grid interface users in the reversed trustworthiness order did neither fixate longer on the most trustworthy search results than list interface users, nor did they select the least trustworthy search results less often than list interface users. These results indicate that with the grid interface laypersons still engage in evaluations of information quality only to a moderate extent. This might be explained by the fact that the cognitive resources of searchers with low prior domain knowledge to engage in profound source evaluations are quite limited, as they have to invest much effort in comprehending the content (cf. Bråten, Strømsø, & Salmerón, 2011). Further

evidence for this conclusion is given by the results on the number of verbal utterances.

In line with Hypothesis 4a, grid interface users verbally reflected more on the *general quality* of search results (e.g., "the search result is good"). For the more specific evaluation criterion *type of source* the effect of the grid interface stimulating more evaluative utterances on the type of source, however, was moderated by participants' age. Older students showed more utterances related to the type of source in the grid interface than in the list interface, whereas the number of utterances of younger students was not increased by the grid interface. Therefore, Hypothesis 4a was at least partly confirmed. With regard to the number of verbal utterances related to *credibility*, which was defined as the most specific of the quality-related evaluation criteria in the present study, however, there were no differences between search interfaces. Furthermore, it should be noted that in particular verbal utterances related to the general quality and credibility were very rare. With regard to number of verbal utterances related to the *position of the search results* in the SERPs, the reduction of such utterances in the grid interface as compared to the list interface was only found for students with high beliefs that the Web contains correct knowledge. Students with doubts about this issue, in both interfaces seldom reflected on the search result position.

Finally, the expected positive effect of the grid interface on participants' search outcome with regard to the medical problem in the reversed trustworthiness order (Hypothesis 5a), was confirmed only for students with high beliefs that the Web contains correct knowledge, but not for students with doubts (i.e., low beliefs) about this issue. When the top search results were of low trustworthiness, students with high beliefs that the Web contains correct knowledge who used the grid interface listed significantly more arguments from the three most trustworthy sources than those who used the list interface. In contrast, students with doubts that the Web contains correct knowledge performed equally well in the two interfaces with regard to the number of arguments from the three most trustworthy sources they listed.

6.4.2 Relationships between Internet-specific epistemic beliefs and the evaluation of information quality

With regard to participants' Internet-specific epistemic beliefs, the hypothesis that doubts that the Web contains correct knowledge would result in more thorough evaluations of the search results (Hypothesis 1b) was confirmed by eye-tracking data of this study. During their Web search, the more doubts students had that the Web contains correct knowledge, the longer they attended to the search results on the SERPs, indicating an increased time spent on deciding which Web pages to access for further inspection. Furthermore, when the top search results were of low trustworthiness (i.e., reversed trustworthiness order), students with doubts that the Web contains correct knowledge verbally reflected more on the *type of source* of search results than when the top search results were of high trustworthiness. Thus, searchers with doubts about the quality of Web information seem to verbally reflect on the type of source only when encountering dubious information sources among the top search results, but not when encountering high-quality information sources in the top positions, which is plausible. In the latter case, instead, they verbally reflected more on the general quality of information than when the top search results were of low trustworthiness. The number of respective verbal utterances of students with high beliefs that the Web contains correct knowledge, in contrast, were not affected by the trustworthiness order in which search results were presented. Besides, in the present study participants' epistemic beliefs did neither relate to the number of verbal utterances related to the *credibility* of information sources, nor to participants' selection behavior on the SERPs.

Moreover, contrary to the expectations, epistemic beliefs did not moderate the effects of the search interface with regard to students' search and evaluation processes during Web search, except for utterances related to the position of the search results in the SERPs. The abovementioned findings indicate that also in a standard list interface low-knowledge searchers' epistemic doubts with regard to the quality of Web information were activated, in that those students with doubts that the Web contains correct knowledge engaged in increased source evaluations, in particular when they encountered dubious information sources among the top results. Yet, as reported in the section above, there was an interaction between interface, trustworthiness order, and epistemic beliefs regarding participants' search outcome, or more precisely, regarding the number of arguments listed from the three most

trustworthy sources. However, contrary to the assumed combined effect of a) a high level of doubts with regard to the quality of Web information and of b) the presentation of search results in a grid interface, on the quality of participants' search outcome (Hypothesis 5b), the grid interface compensated for a low level of doubts (i.e., high beliefs) about this issue. When the top search results were of low trustworthiness, students with high beliefs that the Web contains correct knowledge who used the grid interface listed significantly more arguments from the three most trustworthy sources than those who used the list interface. In other words, whereas in the list interface less doubtful searchers showed detrimental search outcomes in case that the most trustworthy search results were the lowest-ranked ones in the SERPs, in the grid interface they yielded the same level of search outcomes as searchers with a higher level of doubts with regard to the quality of Web information.

6.4.3 Conclusions

In summary, across different types of processing data as well as for the search outcome the results of the present study provide evidence that the search interface (i.e., the interface design of SERPs) and Internet-specific epistemic beliefs, or a combination of both factors, respectively, influence university students' source evaluations during Web search on a complex and unknown medical issue. The presentation of Web search results by means of a grid interface seems to support a more free exploration of the search results and to refrain searchers from focusing on only the first few results presented by the search engine. However, as compared to the list interface, the grid interface led searchers neither to attend longer to the most trustworthy search results, nor to select the least trustworthy search results less often or to verbally reflect more on the credibility of information sources, respectively. These results indicate that with the grid interface low-knowledge searchers still engage in evaluations of information quality only to a moderate extent.

Furthermore, verbal data obtained from cued retrospective reportings suggests that searchers' engagement in profound source evaluations regarding the type or credibility of information sources also depends on students' age. With regard to verbal evaluations related to the type of source in the present study only older students showed an increase in utterances in the grid interface as compared to the list interface. Younger students' utterances related to the type of source, in contrast, were not increased by means of the grid interface. Furthermore, the older the students, the

more they verbally reflected on the credibility of search results, that is, on the perceived motives or expertise of information sources, though, irrespective of the search interface or the trustworthiness order. A potential reason for the effects of age might be that more profound evaluations of information quality require a certain amount of source knowledge which might grow with science-related search experience the students gain during their academic career. Moreover, the lacking difference between interfaces with respect to specific credibility-related utterances might be because the search result descriptions – of both the list interface and the grid interface – provided only sparse and non-salient quality-related information. Therefore, Study 3 (Chapter 7) examined the effects of a search interface that aims at both reducing the salience of the ranking and increasing the salience of quality-related source information in the SERPs.

Furthermore, the results of the present study are in line with the assumption by Bråten et al. (2005) that students believing that the Web is a reliable resource that contains correct knowledge about study-related contents might not realize the epistemic challenges involved in managing the wealth of information and evaluating the different types of information sources. When evaluating the search results provided in the SERPs, students with high beliefs that the Web contains correct study-related knowledge visually inspected the search results less thoroughly and verbally reflected less on the type of source than students who were more skeptical about this issue. Furthermore, when the top search results in a standard Google-like list interface were of low trustworthiness, students with high beliefs that the Web contains correct knowledge showed poorer search outcomes than students who were more skeptical about this issue.

With regard to the references to source information in participants' argumentations, however, all participants in the present study hardly referred to source information at all in their argumentations. A potential reason for this might be that the given task of only listing the pro and con arguments instead of writing a coherent argumentative summary might not have afforded to refer to the sources of the respective arguments. Therefore, in Study 3 of this dissertation (see Chapter 7) an argumentative summary task was chosen instead of a simple listing of arguments.

Finally, it should be acknowledged that the present study comes with certain limitations. First, the trustworthiness order used in the present study was based on

laypersons' rankings, which, of course, could be erroneous. Moreover, to perform the ranking task, participants of the pilot study were only provided with the search results descriptions. It can be assumed that when provided with the Web pages their rankings would have been different. However, the reason for providing participants of the pilot study only with the search results descriptions was that the main focus of the present study was on searchers' spontaneous evaluation and selection behavior on SERPs. Second, the heterogeneous viewing sequences of grid interface users might have at least partly resulted from an initial orienting response to the novel layout. Thus, future research should examine the effects of a search interface presenting search results in a grid during a longer period of time. Moreover, in the case that searchers knew the ordering of the search results in a grid interface, differences between the interfaces might be reduced. Finally, the results concerning epistemic beliefs are only correlative in nature not warranting conclusions about causality.

Notwithstanding these limitations, this study provides novel insights into the impact of the search interface as well as Internet-specific epistemic beliefs on laypersons' source evaluations during Web search on a complex medical issue. In order to further examine these effects, the following chapter presents Study 3 of this dissertation, which investigated the impact of Internet-specific epistemic beliefs and a search interface that is augmented by quality-related information.

7 Study 3: The role of Internet-specific epistemic beliefs and of source cues in search interfaces in evaluating information quality during Web search on medical and health information

7.1 Research questions and hypotheses

Study 2 of this dissertation has shown across different types of processing data that the presentation of search results in a grid interface with a reduced salience of the ranking of the search results increased university students' source evaluations during Web search on a complex and unknown medical issue. Likewise, epistemic doubts that the Web contains correct knowledge about study-related contents were related to increased source evaluations. Moreover, with regard to students' search outcomes results indicated that the grid interface compensated for a low level of doubts with regard to the quality of Web information, such that with the grid interface irrespective of their epistemic beliefs all participants achieved equal search outcomes that were as good as or better than with the list interface. However, the findings of Study 2 also indicated that the grid interface did not result in a decreased selection of least trustworthy, that is, rather dubious, search results, nor did it succeed in increasing searchers' evaluative utterances regarding the credibility of information sources. Similarly, students' epistemic beliefs did not relate to their selection decisions or to their verbal reflections on credibility. To conclude, even with the grid interface and/or a high level of doubts about the quality of Web information participants in Study 2 engaged in evaluations of information quality only to a moderate extent. A potential reason for these findings is that search result descriptions in both the list interface and the grid interface provided only sparse and non-salient quality-related information, thus making it difficult for searchers with low prior domain knowledge to engage in profound source evaluations, for instance, with regard to the credibility of information sources. Therefore, a central goal of the present study, which in line with Study 2 focused on the first evaluation phase of evaluating search results, was to investigate the effects of a search interface that aims

at both reducing the salience of the ranking and simultaneously increasing the salience of quality-related source information in the SERPs on low-knowledge searchers' evaluations of information quality. Furthermore, the study aimed at investigating whether the provision of quality-related source information in the SERPs would compensate for a low level of doubts about the quality of information on the Web (cf. Kammerer et al., 2009), thereby particularly helping searchers with less appropriate epistemic beliefs to access high-quality information sources.

To test the assumption that a search interface with a reduced salience of the ranking of search results and an increased salience of quality-related source information would facilitate evaluations of information quality on SERPs as compared to a standard *list interface*, in the present study an experimental SERP interface that presents search results in a tabular format with labeled columns, in which search results were grouped according to specific source categories was constructed. This so-called *tabular interface* intended to combine three different design approaches, namely (1) using a grid format as in Study 2, (2) presenting additional source cues (e.g., Kammerer et al., 2009; Schwarz & Morris, 2011; Winter et al., 2010; see Chapter 5.2.2), and (3) grouping the search results according to categories (Dumais et al., 2001) in order to provide a type of "representational guidance" (Stadtler & Bromme, 2008; Suthers & Hundhausen, 2003; see also Chapter 5.2.2). This combination of design features was expected to results in appropriate affordances for searchers to engage in processes of source evaluation.

The source categories according to which the search results were grouped in the present study referred to the "objectivity" dimension of information quality (cf. Tate, 2010; see Chapter 2.2), that is, the extent to which information is presented without distortion by personal feelings or other biases (e.g., commercial interests). Based on work by Finn and Kushmerick (2006) who developed an automatic classifier that distinguishes Web news articles according to the categories "objective" (representing factual reports) and "subjective" (representing opinions), in the present study Web pages were (manually) classified according to the source categories "objective information" and "subjective information". Furthermore, a third category, namely "commercial information" (representing the purpose to promote or sell) was added.

It was hypothesized that such a tabular interface, which externally represents knowledge about sources in a structured way, would guide students with low prior

domain knowledge more in their Web search and would support more their evaluations of information quality than a standard list interface. As a consequence, the use of the tabular interface during Web search should also result in better search outcomes than the list interface. Moreover, the tabular interface was assumed to compensate for a low level of doubts with regard to the quality of Web information, such that students with *high* beliefs that the Web contains correct knowledge should benefit most from the tabular interface by becoming more aware about the diversity of information sources or potential biases than in the list interface. Accordingly, these students were assumed to show an increased engagement in source evaluations when using the tabular interface as compared to the list interface. For students with doubts that the Web contains correct knowledge a smaller effect of interface on their engagement in source evaluations was expected, as in line with findings from Study 2 they were assumed to also engage in source evaluations when using the list interface. Figure 31 graphically represents this assumed interaction of search interface and epistemic beliefs on participants' engagement in source evaluations during Web search.
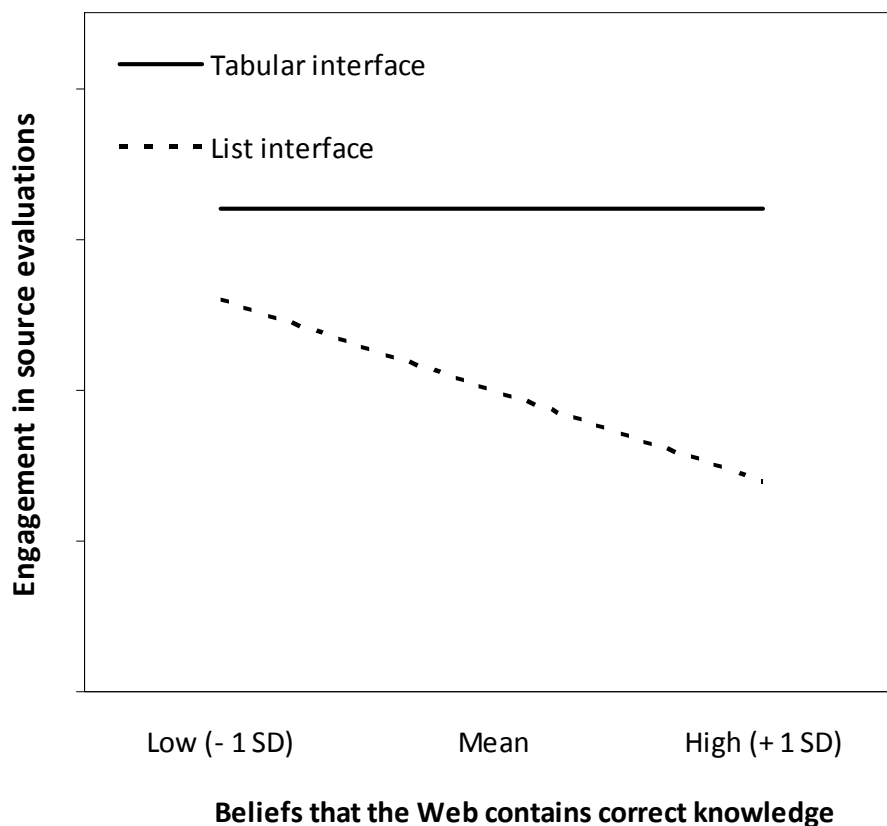
Figure 31. Assumed interaction pattern between search interface and epistemic beliefs on the engagement in source evaluations.

Evaluation processes on SERPs in the present experimental study again were assessed through a combination of eye tracking methodology, log file data (i.e., mouse clicks), and cued retrospective verbal protocols (cf. Study 2). The solution to the information problem was analyzed by asking participants to write an argumentative essay about the medical issue.

According to these measures, the assumption that both a high level of doubts that the Web contains correct knowledge and a search interface with an increased salience of quality-related source information would facilitate evaluations of information quality on SERPs and that such a search interface furthermore would compensate for a low level of doubts with regard to the quality of Web information, resulted in the following more specific hypotheses.

7.1.1 Eye-tracking data

In line with findings from Study 2, it was assumed that the more participants doubted (i.e., the lower their beliefs) that the Web contains correct knowledge, the longer they would attend to search results on the SERPs (Hypothesis 1a). Furthermore, it was hypothesized that the tabular interface supported participants in their source evaluations during Web search, such that in the tabular interface high-quality (i.e., objective) search results would be attended longer and lower-quality, potentially biased search results (i.e., in particular commercial sources) would be attended shorter than in the list interface (Hypothesis 1b). Moreover, the increased attention to high-quality search results and the reduced attention to lower-quality search results in the tabular interface as compared to the list interface was expected to be larger for participants with high beliefs that the Web contains correct knowledge than for participants with doubts about this issue (Hypothesis 1c).

7.1.2 Search result selection

Bases on the results from Study 2, no main effect of epistemic beliefs on the selection of search results was assumed (Hypothesis 2a). However, in line with the assumptions for the eye-tracking data, it was hypothesized that in the tabular interface more high-quality (i.e., objective) search results and less lower-quality, potentially biased search results (i.e., in particular commercial sources) would be selected than in the list interface (Hypothesis 2b). In addition, based on the assumption that the source categories would provide representational guidance, high-

quality (i.e., objective) search results were assumed to be selected more often as the first ones than in the list interface (Hypothesis 2c). Moreover, the increased and primary selection of high-quality search results and the reduced selection of lower-quality search results in the tabular interface as compared to the list interface was expected to be larger for participants with high beliefs that the Web contains correct knowledge than for participants with doubts about this issue (Hypothesis 2d).

7.1.3 Verbal utterances

In line with findings from Study 2, it was assumed that the more participants doubted (i.e., the lower their beliefs) that the Web contains correct knowledge, the more they would verbally reflect on the information quality (e.g., referring to the type of source) of the search results (Hypothesis 3a). Furthermore, it was hypothesized that with the tabular interface because of the reduced salience of the ranking and the increased salience of quality cues, the number of verbal utterances addressing the position of search results on a SERP would be decreased and the number of quality-related utterances (e.g., about the type or credibility of information sources) would be increased as compared to the list interface (Hypothesis 3b). Moreover, the decrease of utterances addressing the position of search results and the increase of quality-related utterances in the tabular interface as compared to the list interface was assumed to be larger for participants with high beliefs that the Web contains correct knowledge than for participants with doubts about this issue (Hypothesis 3c). Only with regard to profound verbal evaluations related to the credibility of information sources the increase was expected to be larger for participants with doubts (i.e., low beliefs) that the Web contains correct knowledge than for participants with high beliefs about this issue (Hypothesis 3d).

7.1.4 Quality of information problem solving

Bases on the results from Study 2, no main effect of epistemic beliefs on the quality of participants solution to the information problem (i.e., their search outcome) was assumed (Hypothesis 4a). However, if users of the tabular interface selected more high-quality search results and less lower-quality search results than users of the list interface, then in the tabular interface the quality of students' search outcome should be improved as well. It was assumed that users of the tabular interface would bring in more arguments presented by objective sources in their argumentative summaries

about the medical issue than list interface users (Hypothesis 4b). Moreover, the advantage of the tabular interface over the list interface with regard to the quality of participants' search outcome was expected to be larger for participants with high beliefs that the Web contains correct knowledge than for participants with doubts about this issue (Hypothesis 4c).

7.2   Method

7.2.1   Participants

Participants were 58 university freshmen (10 male, 48 female; $M = 20.52$ years, $SD = 1.59$; range 19 – 28 years) from different majors at the University of Tuebingen, Germany; participation was rewarded with either course credit or payment. Pharmacy and medical students were excluded from the study. To ensure that prior domain knowledge was low, participants' prior knowledge on the content domain of the study (Bechterew's disease and other rheumatic diseases) was assessed (see section 7.2.3.5 for details), revealing scores from 1.00 to 2.50 ($M = 1.24$, $SD = 0.32$) on a scale from 1 (low) to 5 (high). Participants had normal or corrected to normal vision. All participants reported to use Google as their primary search engine.

7.2.2   Experimental design

The experiment was based on a three-factorial mixed-model design. As a first factor the *search interface* was varied between subjects. Participants either received SERPs in the form of a standard *list interface* (see Figure 32) with nine search results being listed from top to bottom on each SERP, or in the form of a *tabular interface* (see Figure 33) with the same nine search results being grouped in three columns, labeled as "objective information" (e.g., a Web page from the department of health, or from a reputable medical magazine), "subjective information" (e.g., a forum page from a support group), or "commercial information" (e.g., a Web page provided by a drug company or a health farm). A pilot study with 10 university students being asked to list characteristics and examples of objective, subjective, and commercial Web information confirmed that this classification was comprehensible. All students listed correct characteristics and examples for the three categories. To avoid order effects, for each participant the search results in the list or in the columns were presented in random order and, furthermore, in the tabular interface the order of the three columns

was permuted. Participants were randomly assigned to the two search interface conditions, with 28 participants serving in each condition.
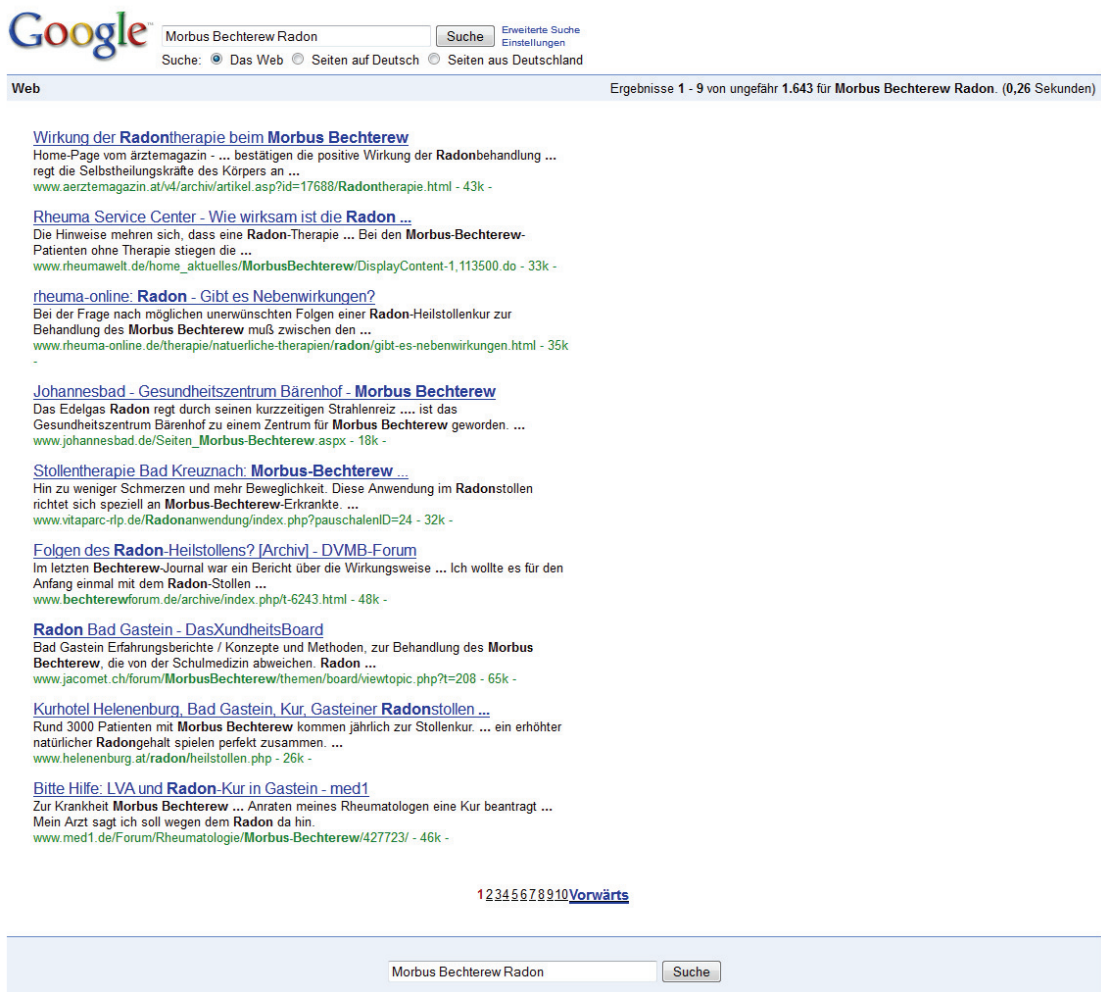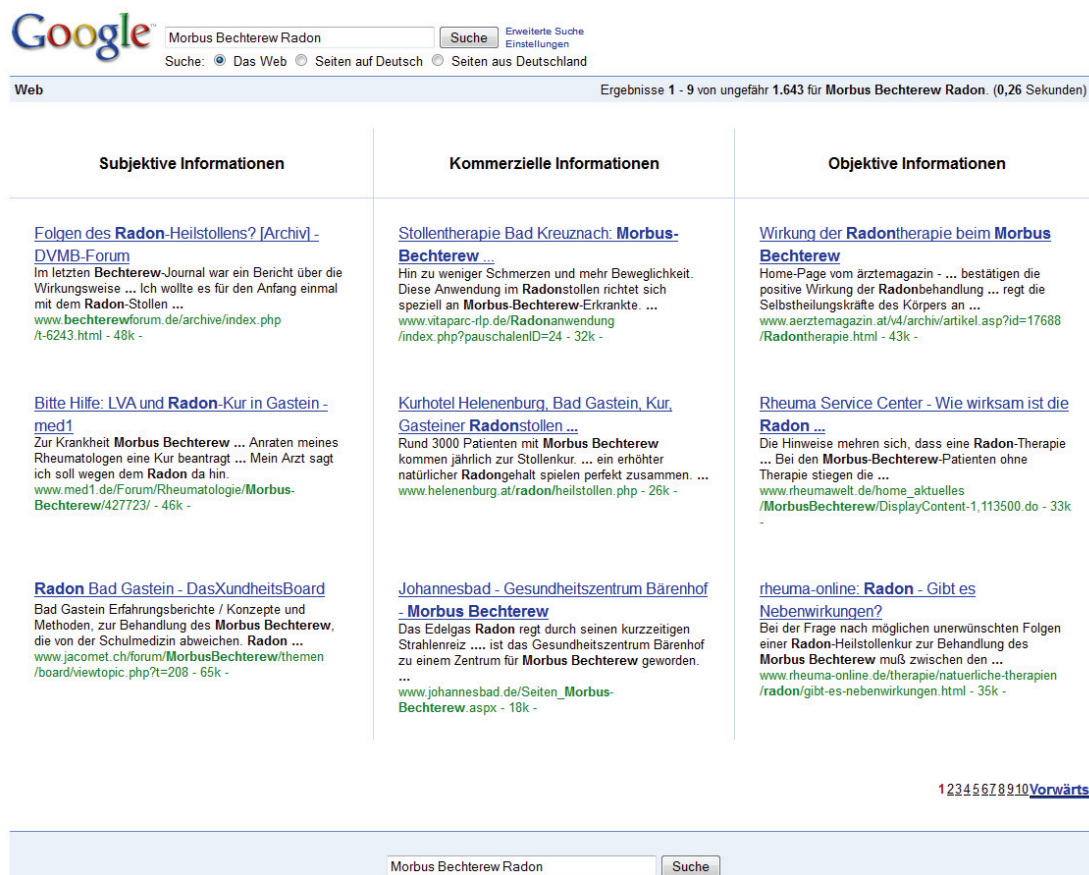

Figure 32. List interface.

Figure 33. Tabular interface.

Search result category, that is, to which of the three categories (objective, subjective, or commercial information) a search result or the corresponding Web page belonged, was considered as a second within-subject factor.

As a third factor participants' *epistemic beliefs* about whether or not the Web contains correct knowledge (see section 7.2.3.4 for details) were assessed and used as a continuous between-subject factor.

7.2.3 Materials and apparatus

7.2.3.1 Task

As in Study 2 the task used in the experiment was to seek information on the WWW about two competing therapies ('radon therapy' and 'infliximab therapy') for Bechterew's disease (i.e., a chronic inflammatory rheumatic disease affecting the spine) in order to give informed advice about the therapies to a fictitious friend who was recently diagnosed with the disease. As in Study 2, to ensure that participants preselected and focused on some of the information sources – which would reflect a

real search situation on the WWW – rather than reading all available information, the total search time was limited to eight minutes.

7.2.3.2 Web materials

To complete the experimental task, as in Study 2 participants were provided with two prefabricated SERPs (one for each therapy) accessible by means of a start Web page (for sample screenshots see Appendix G). Each of the two SERPs reflected the heterogeneity of information sources on the Web, including three objective search results (e.g., scholarly or informational Web pages providing neutral, fact-based information), three subjective search results (i.e., forum Web pages providing personal opinions and experiences), and three commercial search results (i.e., Web pages from health farms or drug companies promoting their treatments). For an appropriate assignment of the search results to one of these three categories, three search results and corresponding Web pages from the materials of Study 2 were modified or replaced. Apart from the experimental manipulation of the interface (list vs. tabular interface), as in Study 2, the SERPs were displayed in Google style (cf. logo, font style and colors of search results, etc.). However, sponsored links (i.e., ads) and the hyperlinks "in cache" and "similar pages" were not included on the SERPs. Search result links that have already been selected were marked in purple color, whereas not yet selected links were displayed in blue. Web materials were presented with Microsoft Internet Explorer 7.

7.2.3.3 Apparatus

The setup to record participants eye movements, mouse clicks, and retrospective verbal reports was the same as in Study 2. Furthermore, as in Study 2, the recorded gaze data were mathematically corrected posthoc for potential systematic offsets by means of an interpolation algorithm (Kammerer, 2010).

7.2.3.4 Internet-specific epistemic beliefs measure

To assess the extent to which participants believed that the Web contains correct knowledge, a translated and adapted version of the scale "certainty and source of knowledge" (8 items; Cronbach's α = .72) of the Internet-Specific Epistemological Questionnaire (ISEQ, Strømsø & Bråten, 2010) was used. In contrast to Study 2, in the present study the eight items were formulated with regard to Web-based knowledge in general instead of study-related knowledge, because participants were

all university freshmen, who can be assumed to have only little experience with study-related knowledge on the Web. The items had to be rated on 5-point Likert-type response scales ranging from 1 (*totally disagree*) to 5 (*totally agree*). High scores represented the belief that the Web is a reliable resource that contains correct knowledge. Low scores represented doubts about this issue. Sample items state that the Web contains accurate knowledge, most of what is true about a particular topic, or expert statements that help to resolve difficult problems (for the full list of items as well as means and standard deviations of the items, see Appendix H). Thus, the items used in the present study measured slightly different epistemic beliefs than in Study 2. In contrast to the items used in Study 2 that specifically addressed Web-based knowledge about study-related contents, the items used in the present study addressed knowledge on the Web in general. Thus, when answering the statements, participants might have different domains in mind (e.g., sports, politics, health, finance, news, etc.) for which they use the Web as information resource.

To avoid Web search performance being affected by thoughts provoked by the questionnaire, the questionnaire was administered online about one week after the experiment. Scores on the epistemic beliefs scale were normally distributed ($W = 0.98$, $p = .47$) with a mean of 3.26 ($SD = 0.56$) and a range from 1.63 to 4.38. There were no differences between the two experimental conditions regarding participants' certainty and source beliefs, $t(56) = -0.93$, $p = .36$ (see Table 13 for means and standard errors).

7.2.3.5 Control variables

Demographics (gender, age), computer- and Web search experience and skills (3 items, see Appendix C; Cronbach's $\alpha = .75$), and prior knowledge on Bechterew's disease or other rheumatic diseases and respective therapies (10 items, see Appendix F, Cronbach's $\alpha = .64$) were assessed as control variables (see Table 13 for means and standard errors). Except for gender and age, items had to be rated on five-point Likert-type response scales ranging from 1 (*totally disagree* or *very low*) to 5 (*totally agree* or *very high*). Analyses of the respective data revealed no differences between the two experimental conditions, that is, for gender, $\chi^2(1, N = 58) = 0.00$, $p > .99$, for age, $t(56) = 0.33$, $p = .75$, for computer- and Web search experience and skills, $t(56) = -0.93$, $p = .36$, and for prior domain knowledge, $t(56) = -0.62$, $p = .54$.

Table 13

Means (and Standard Deviations) of Internet-specific Epistemic Beliefs and of Control Variables as a Function of Interface

	List interface (*n = 28)*	Tabular interface (*n = 28*)
Age	20.59 (1.40)	20.44 (1.78)
Gender	5 m, 24 f	5 m, 24 f
Epistemic beliefs 1 (low)– 5 (high)	3.20 (0.57)	3.33 (0.56)
Prior domain knowledge 1 (low) – 5 (high)	1.21 (0.27)	1.27 (0.36)
Computer- and Web search experience and skills 1 (low) – 5 (high)	3.46 (0.63)	3.60 (0.49)

7.2.4 Procedure

Testing procedure was equivalent to the one used in Study 2 except for the fact that in the present experiment participants were given five minutes to write an argumentative summary about the pros and cons of the two therapies, instead of simply listing the arguments.

7.2.5 Data analyses and dependent variables

7.2.5.1 Eye-tracking data

To analyze participants' eye-tracking data, for each of the nine search results on a SERP a polygonal area of interest (AOI) was defined covering the title, excerpt, and URL of a search result. It was determined for all AOIs whether and for which amount of time a participant was looking at this area. Although the shape of the AOIs differed between the two search interfaces (list vs. tabular interface), their size, text content, font style, and font size were exactly the same for both interfaces (see Figures 32 and 33). As no differences were expected between the two SERPs (i.e., the 'radon SERP' and the 'infliximab SERP') eye-tracking data of both SERPs were aggregated in the statistical analyses. As dependent variable the *total fixation time* (in milliseconds) on *objective, subjective, and commercial search results*, that is, the

total time for which participants attended to the six search results of one of the three categories during their eight-minute Web search, was assessed.

7.2.5.2 Log file data (mouse clicks)

With regard to participants' selection data, the search results participants selected from the SERPs during their eight-minute Web search to access a Web page were recorded. As a first dependent variable the *number of objective, subjective, and commercial search results selected* from the SERPs was analyzed. Selection data were aggregated across the two SERPs (i.e., the 'radon SERP' and the 'infliximab SERP'). Accordingly, for each of the three categories (i.e., objective, subjective, and commercial search results) a maximum of six search results could be selected (i.e., re-openings of a Web page were not counted). As a second dependent variable the frequency with which the *first search result selected* from a SERP was an *objective search result* was analyzed.

7.2.5.3 Coding scheme for thinking-aloud protocols

For the analysis of participants' cued retrospective verbalizations obtained while replaying the eye-movement recordings of the two Google SERPs, the coding scheme of Study 2 was used (cf. section 6.2.5.3). Accordingly, as dependent variables the number of verbal utterances related to the four evaluation criteria *search result position*, *general quality*, *type of source*, and *credibility* were analyzed. Verbal utterances expressed across the two SERPs were aggregated in the analyses. Short descriptions of the four evaluation criteria are provided in Table 14. Two raters familiar with the search task and the Web materials as well as with the coding scheme scored 25% of the protocols. Inter-rater reliability computed on this subsample of protocols yielded a Cohen's kappa of $k = .70$. One rater scored the remaining protocols and only the coding of this rater was used for further data analyses. Coding was realized with the software tool MEPA 4.10 (Erkens, 2005).

Table 14.

Coding Scheme for Thinking-aloud Protocols

Evaluation criteria	Short description
Search result position	The search result is selected because it is the first, the highest ranked, the topmost search result.
General quality	The search result or the corresponding web page is (or is not) good.
Type of source	The type of source, e.g., a forum, a drug company, the department of health., is mentioned without an explicit credibility rating. In the tabular interface this category also included mere references to objective, subjective, or commercial search results.
Credibility	The source is (or is not) expected to provide trustworthy, reliable, neutral, official information; information provided by an expert.

7.2.5.4 Quality of information problem solving

As the product of the search task, participants' argumentative summaries were analyzed with regard to the arguments expressed in favor and against a therapy. Specifically, the *number of arguments presented by objective Web pages* ("objective arguments") was coded using a detailed coding rubric. We identified 40 different (but partially overlapping) statements which could be used as objective arguments in favor or against the therapies. Examples are "a placebo-controlled study confirmed the anti-inflammatory effect of the radon therapy", "radon stabilizes the immune system", "infliximab therapy increases the risk of infections", or "it is scientifically confirmed that infliximab reduces disease activity". It has to be noted, though, that these statements were partly presented in subjective or commercial Web pages as well. In addition, for exploratory analyses it was analyzed whether or not participants referred to the source of information in their argumentative summaries. Two raters familiar with the search task and the Web materials as well as with the coding scheme coded the summaries of 15 participants (25.86%). Inter-rater reliability computed on the number of objective arguments coded in the summaries of this subsample yielded an intraclass-correlation of $r = .87$. Disagreements were resolved through discussion between the raters. One rater coded the remaining summaries.

7.3 Results

An alpha level of .05 was used for the statistical tests reported. Effects were considered marginally significant when p-values were between .05 and .10. Epistemic beliefs were used as a covariate (z-scored) in all analyses. Furthermore, an interaction term between epistemic beliefs and interface was included in the statistical analyses. As in Study 1 and 2, significant interaction effects were further examined and graphed using the procedures outlined by Aiken and West (1991). Specifically, simple comparisons were computed for low epistemic belief scores (defined as one standard deviation below the sample mean; $M - 1\ SD$) and for high epistemic belief scores (defined as one standard deviation above the sample mean; $M + 1\ SD$), using moderated regression analyses. All means and standard errors of the dependent measures reported in the following are corrected for the influence of epistemic beliefs.

7.3.1 Eye-tracking data

During the eight-minute Web search participants on average visually inspected 86.30% ($SE = 1.74$) of the 18 search results. Means and standard errors of the total fixation time on search results as a function of interface and search result category are shown in Figure 34.
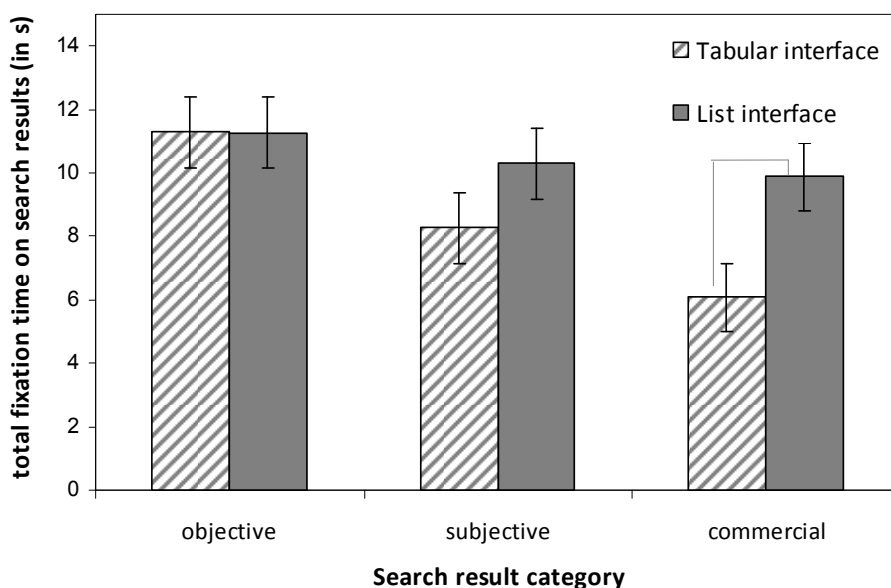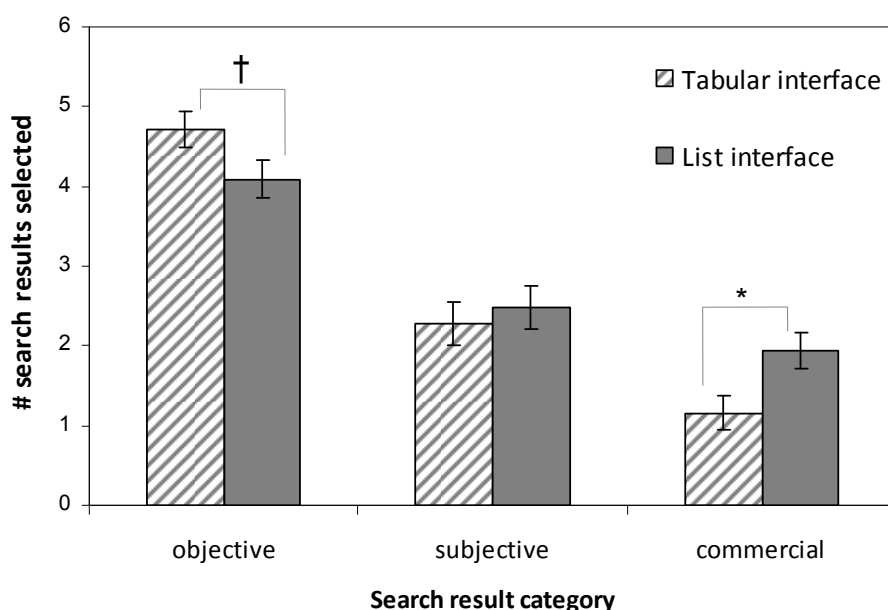

Figure 34. Means (and standard errors) of the total fixation time on search results as a function of interface and search result category. * $p \leq .05$

An ANCOVA with interface (list vs. tabular) as between-subject factor, search result category (objective, subjective, commercial) as within subject-factor, and epistemic beliefs as covariate on the *total fixation time on search results* showed no significant main effect of interface ($F(1, 54) = 2.10$, $p = .15$), but of search result category ($F(2, 108) = 10.28$, $p < .001$, partial $\eta^2 = .16$). However, this main effect has to be interpreted in the light of a significant interaction between interface and search result category on the total time for which search results were fixated ($F(2, 108) = 3.48$, $p = .03$, partial $\eta^2 = .06$). Bonferroni posthoc tests revealed that in the tabular interface commercial search results were visually inspected significantly shorter than in the list interface ($p = .02$). Participants' dwell times on objective and subjective search results, in contrast, did not differ between the two interfaces ($p = .99$ and $p = .20$, respectively). Furthermore, the posthoc tests showed that with the list interface (due to the randomized ordering of the search results for each participant) total fixation times on search results did not differ between the three categories (all $ps > .60$). In contrast, with the tabular interface search results of the objective category were visually inspected significantly longer than both subjective search results ($p = .02$) and commercial search results ($p < .001$). Commercial search results were visually inspected marginally significantly shorter than subjective search results ($p = .07$). Besides, there was no significant main effect of epistemic beliefs on the total fixation time on search results ($F(1, 54) = 1.25$, $p = .27$), nor interaction effects between epistemic beliefs and interface ($F < 1$), epistemic beliefs and search result category ($F(2, 108) = 1.26$, $p = .29$), or a three-way interaction between the factors ($F < 1$).

However, in exploratory analyses the *average fixation time on a non-selected search result*, that is, the total time for which participants fixated search results on the SERPs that they did not select during Web search, divided by the number of non-selected search results. An ANCOVA with interface (list vs. tabular) as between-subject factor and epistemic beliefs as covariate showed a marginally significant main effect of interface ($F(1, 54) = 3.08$, $p = .09$, partial $\eta^2 = .05$). In the list interface a non-selected search result was fixated for $M = 1.21$ s ($SE = 0.14$), whereas in the tabular interface the average fixation time was only $M = 0.87$ s ($SE = 0.14$). Second, there was a marginally significant main effect of epistemic beliefs ($F(1, 54) = 3.08$, $p = .09$, partial $\eta^2 = .05$) showing a positive relationship ($\beta = .23$, $p = .09$) between beliefs that the Web contains correct knowledge and the average time spent fixating a

search result that was not selected during Web search. The interaction between epistemic beliefs and interface, however, was not significant ($F < 1$).

7.3.2 Search result selection

During the eight-minute Web search participants selected 46.32% ($SE = 1.45$) of the 18 search results on average. Means and standard errors of the number of search results selected as a function of interface and search result category are shown in Figure 35.

Figure 35. Means (and standard errors) of the number of search results selected as a function of interface and search result category. * $p \leq .05$, † $p \leq .10$

An ANCOVA with interface (list vs. tabular) as between-subject factor, search result category (objective, subjective, commercial) as within subject-factor, and epistemic beliefs as covariate showed no significant main effect of interface or of epistemic beliefs on the number of search results selected (both Fs < 1). However, there was a significant main effect of search result category ($F(2, 108) = 67.56$, $p < .001$, partial $\eta^2 = .56$) and an interaction between interface and search result category ($F(2, 108) = 3.93$, $p = .02$, partial $\eta^2 = .07$). Bonferroni posthoc tests showed that the number of selected search results differed significantly between the three categories (all ps < .01), with objective search results being selected most often ($M = 4.41$. $SE = 0.16$), commercial search results least often ($M = 1.55$. $SE = 0.16$) and subjective search

157

results in between ($M = 2.38$. $SE = 0.19$). Moreover, in the tabular interface objective search results were selected marginally significantly more often ($p = .06$) and commercial search results significantly less often ($p = .02$) than in the list interface.

In addition, there were significant interactions between interface and epistemic beliefs ($F(1, 54) = 8.90$, $p < .01$, partial $\eta^2 = .14$) and between search result category and epistemic beliefs ($F(2, 108) = 3.90$, $p = .02$, partial $\eta^2 = .07$). With regard to the interaction between interface and epistemic beliefs, simple comparisons conducted according the procedure outlined by Aiken and West (1991) revealed that students with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) selected significantly less search results in the tabular interface than in the list interface ($\Delta M = -1.91$, $SE = 0.74$, $t(54) = -2.58$, $\beta = -.46$, $p = .01$). In contrast, for students with low beliefs that the Web contains correct knowledge ($M - 1\ SD$) there was a trend that in the tabular interface they selected more search results than in the list interface ($\Delta M = 1.23$, $SE = 0.74$, $t(54) = 1.66$, $\beta = .30$, $p = .10$). Figure 36 represents this interaction graphically.
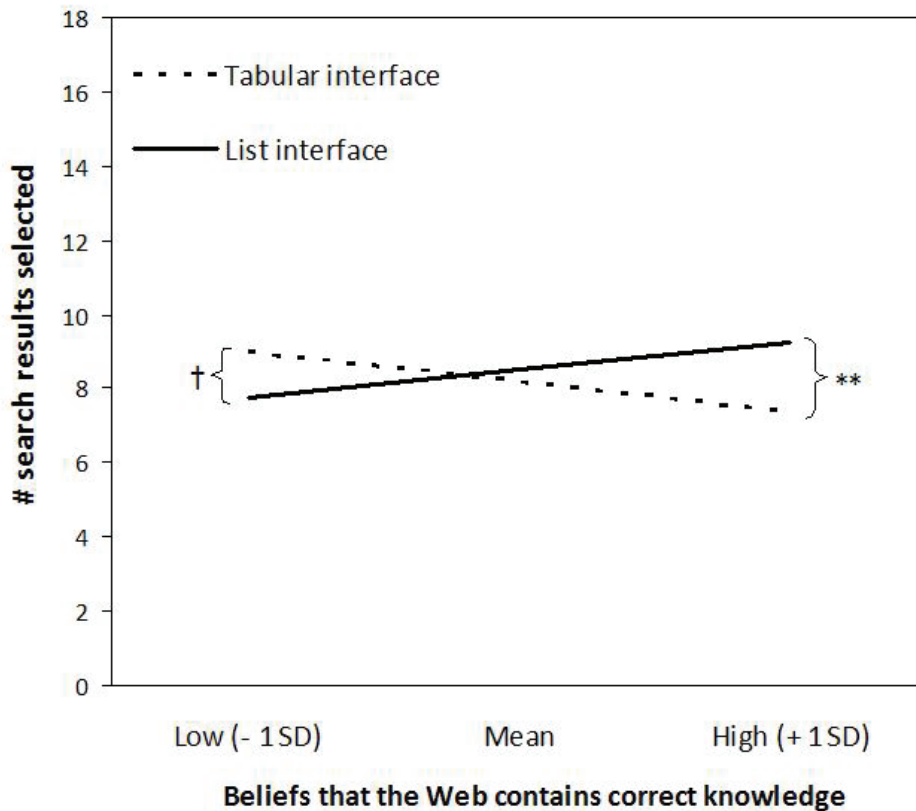
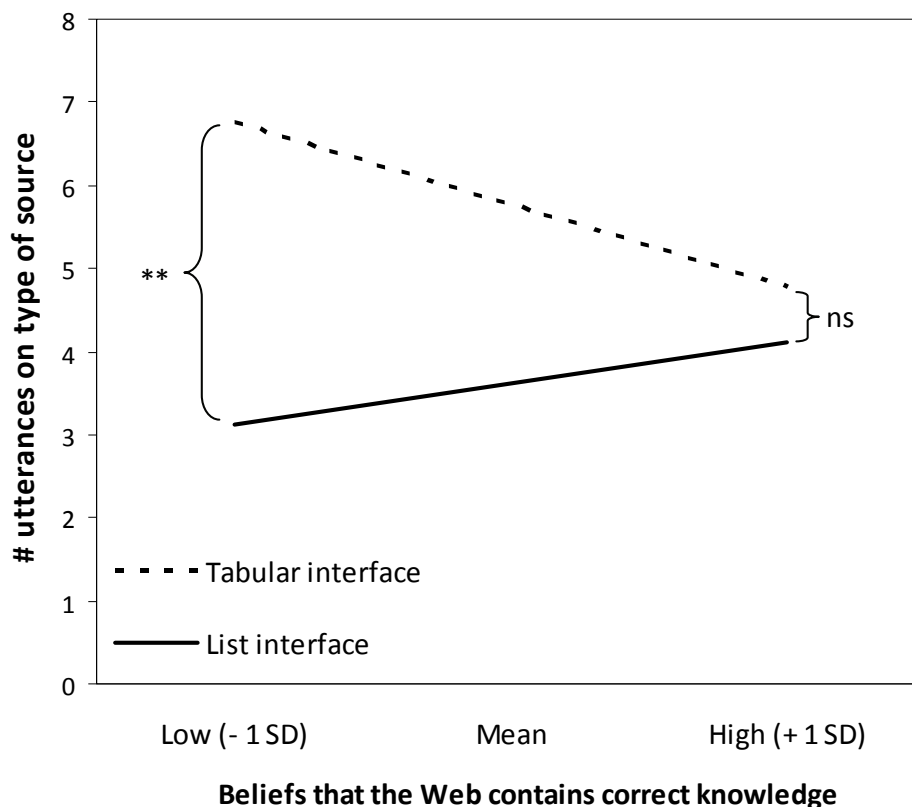*Figure 36.* Interaction of interface and epistemic beliefs with regard to the number of search results selected. ** $p \leq .01$, † $p \leq .10$

The interaction effect between epistemic beliefs and search result category can be traced back to a significant positive relationship between beliefs that the Web contains correct knowledge and the selection of objective search results (β = .28, p = .03), whereas for subjective (β = -.21, p = .13) and commercial search results (β = -.05, p = .70) no significant relationships were revealed. Besides, there was no three-way interaction between interface, search result category, and epistemic beliefs ($F(2, 108) = 2.26$, p = .11). Besides, there was no three-way interaction between interface, search result category, and epistemic beliefs ($F(2, 108) = 2.26$, p = .11).

Finally, regarding the frequency with which the first search result selected from a SERP was an objective search result, an ANCOVA with interface (list vs. tabular) as between-subject factor and epistemic beliefs as covariate showed a significant main effect of interface ($F(1, 54) = 7.61$, p = .01, partial η^2 = .12). In the tabular interface 80.65% (SE = 6.54) of the first selected search results were objective ones, whereas with the list interface it were only 55.14% (SE = 6.54). Second, there was a marginally significant main effect of epistemic beliefs ($F(1, 54) = 3.32$, p = .07, partial η^2 = .06) showing a positive relationship (β = .23, p = .07) between beliefs that the Web contains correct knowledge and the frequency of objective search results selected first. The interaction between epistemic beliefs and interface, however, was not significant ($F(1, 54) = 1.30$, p = .26).

7.3.3 Verbal utterances

Table 15 shows means and standard errors of the number of verbal utterances related to the four evaluation criteria as a function of interface.

Table 15

Means (and Standard Errors) of the Number of Verbal Utterances related to the Four Evaluation Criteria as a Function of Interface

Evaluation criteria	List interface	Tabular interface
Search result position	0.92 (0.17)	0.44 (0.17)
General quality	0.31 (0.15)	0.71 (0.16)
Type of source	3.61 (0.59)	5.78 (0.59)
Credibility	1.42 (0.31)	1.42 (0.31)

ANCOVAs with interface (list vs. tabular) as between-subject factor and epistemic beliefs as covariate were conducted for the number of utterances related to the four evaluation criteria *search result position*, *general quality*, *type of source*, and *credibility*.

7.3.3.1 Number of utterances related to the search result position

With regard to the number of utterances related to the *search result position*, the ANCOVA showed a significant main effect of interface ($F(1, 54) = 4.04$, $p = .05$, partial $\eta^2 = .07$), but not of epistemic beliefs ($F < 1$). The main effect of interface, however, has to be interpreted in the light of a marginally significant interaction between interface and epistemic beliefs ($F(1, 54) = 3.03$, $p = .09$, partial $\eta^2 = .05$).

Simple comparisons revealed that students with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) verbally addressed the position of search results significantly less in the tabular interface than in the list interface ($\Delta M = -0.91$, $SE = 0.34$, $t(54) = -2.58$, $\beta = -.49$, $p = .01$). Students with low beliefs that the Web contains correct knowledge ($M - 1\ SD$), in contrast, did not differ in the number of utterances related to the search result position in the two interface conditions ($\Delta M = -0.06$, $SE = 0.74$, $t(54) = -0.17$, $\beta = -.03$, $p = .86$). Figure 37 represents this interaction graphically.
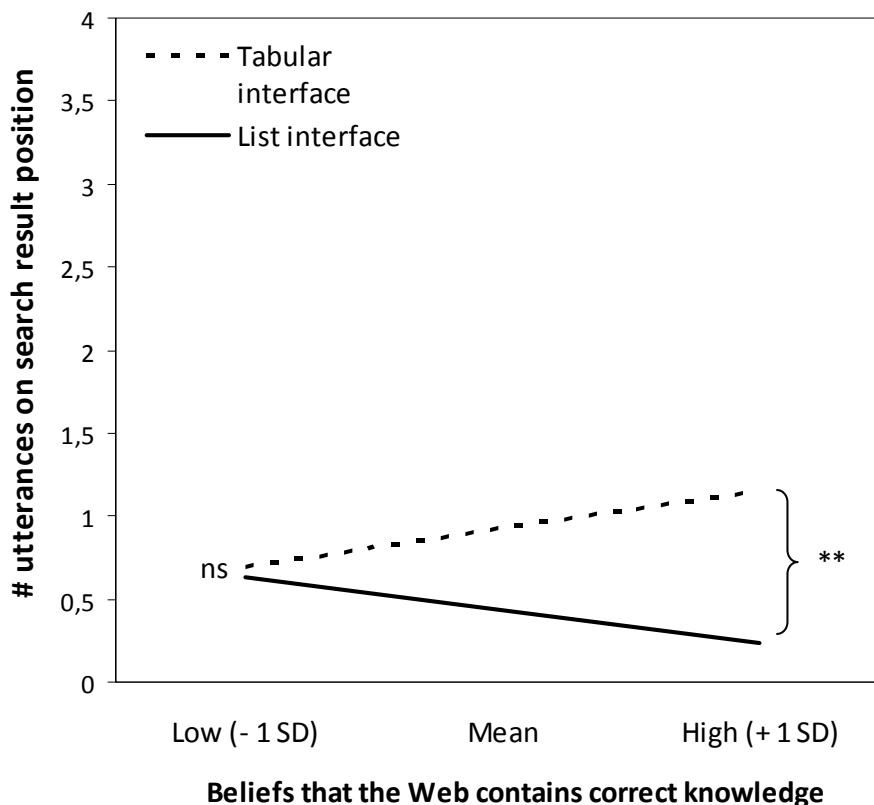
Figure 37. Interaction of trustworthiness order and epistemic beliefs with regard to the number of verbal utterances related to the search result position. ** $p \leq .01$

7.3.3.2 Number of utterances related to the general quality

With regard to the number of utterances related to the *general quality*, the ANCOVA showed a marginally significant main effect of interface ($F(1, 54) = 3.40$, $p = .07$, partial $\eta^2 = .06$). Participants using the tabular interface tended to expressed more utterances than participants using the list interface. Besides, there was neither a significant main effect of epistemic beliefs, nor a significant interaction between interface and epistemic beliefs (both $Fs < 1$).

7.3.3.3 Number of utterances related to the type of source

With regard to the number of utterances related to the *type of source*, the ANCOVA showed a significant main effect of interface ($F(1, 54) = 6.70$, $p = .01$, partial $\eta^2 = .11$), but not of epistemic beliefs ($F < 1$). However, the main effect of interface has to be interpreted in the light of a marginally significant interaction between interface and epistemic beliefs ($F(1, 54) = 3.10$, $p = .08$, partial $\eta^2 = .05$). Simple comparisons revealed that students with low beliefs that the Web contains correct knowledge (*M* -

1 *SD*) verbally reflected on the type of source significantly more in the tabular interface than in the list interface (ΔM = 3.65, *SE* = 1.19, $t(54)$ = 3.07, β = .55, p < .01). Students with high beliefs that the Web contains correct knowledge (M + 1 *SD*), on the contrary, did not differ in the number of verbal utterances related to the type of source in the two interface conditions (ΔM = 0.68, *SE* = 1.19, $t(54)$ = 0.57, β = .10, p = .57). Figure 38 represents this interaction graphically.

Figure 38. Interaction of interface and epistemic beliefs with regard to the number of verbal utterances related to the type of source. ** $p \leq .01$

7.3.3.4 Number of utterances related to credibility

With regard to the number of utterances related to the *credibility* of information sources, the ANCOVA showed neither significant main effects of interface or epistemic beliefs, nor a significant interaction between the two factors (all *F*s < 1).

7.3.4 Quality of information problem solving

With regard to the number of objective arguments in the summaries written by participants, the ANCOVA with interface as between-subject factors and epistemic

beliefs as covariate showed no main effect of interface ($F < 1$), with participants using the tabular interface having included 4.54 ($SE = 0.47$) objective arguments and participants using the list interface having included 4.52 ($SE = 0.47$) objective arguments in their summaries. Yet, there was a marginally significant main effect of epistemic beliefs ($F(1, 54) = 3.52$, $p = .07$, partial $\eta^2 = .06$) and a significant interaction between interface and epistemic beliefs ($F(1, 54) = 6.52$, $p = .01$, partial $\eta^2 = .11$). Simple comparisons conducted according the procedure outlined by Aiken and West (1991) revealed that students with high beliefs that the Web contains correct knowledge ($M + 1\ SD$) included marginally significantly more objective arguments in their summaries when they had used a tabular interface than when they had used the list interface ($\Delta M = 1.79$, $SE = 0.93$, $\beta = .35$, $t(54) = 1.92$, $p = .06$). In contrast, for students with low epistemic beliefs ($M - 1\ SD$) the interface did not significantly influence the number of objective arguments included in their summaries ($\Delta M = -1.19$, $SE = 0.94$, $\beta = -.23$, $t(54) = -1.28$, $p = .21$). Figure 39 represents this interaction graphically.
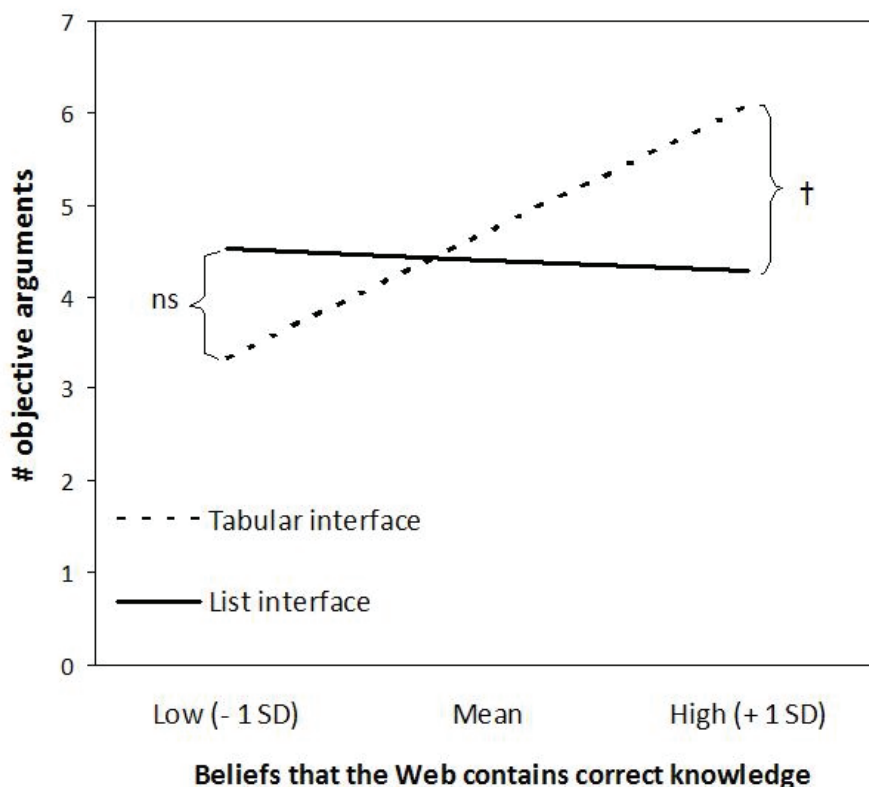
Figure 39. Interaction of interface and epistemic beliefs with regard to the number of objective arguments included in the argumentative summary. † $p \leq .10$

Finally, exploratory analyses with regard to source references (e.g., "according to a placebo-controlled study", "according to information from pharmaceutical industry", "according to personal reports") included in the argumentative summaries showed no differences between interfaces ($\chi^2(1, N = 58) = 0.70, p = .40$). Eight list interface users (27.59%) and 11 tabular interface users (37.93%) referred to the source of information in their argumentative summaries.

7.4 Discussion

The purpose of the present study was to examine the role of both Internet-specific epistemic beliefs and the search interface in facilitating or hampering low-knowledge searchers' evaluations of information quality during Web search on a complex medical issue. Specifically, in line with the results of Study 2 it was assumed that the more doubts searchers had that the Web contains correct knowledge, the more appropriate source evaluations they would engage in on SERPs. It should be noted that because participants in the present study were university freshmen with only few Web experiences with regard to study-related contents, other than in Study 2 participants' epistemic beliefs regarding Web-based knowledge in general were assessed instead of Web-based knowledge about study-related contents. Besides, with regard to the search interface it was hypothesized that an interface with a reduced salience of the ranking of the search results and an increased salience of quality-related source information would facilitate the evaluation of information quality on SERPs as compared to a list interface that imposes a high salience on the ranking of the search results and that provides only sparse and non-salient quality-related source information. Therefore, the present study compared a standard Google-like list interface to a tabular interface with additional source cues (objective information, subjective information, commercial information) according to which the search results on the SERPs were grouped. Moreover, the tabular interface was assumed to compensate for a low epistemic awareness of the varying information quality on the Web, such that students with *high* beliefs that the Web contains correct knowledge should benefit most from the tabular interface by becoming more doubtful with regard to quality of the information provided or with regard to potential biases than in the list interface.

With regard to epistemic beliefs, results of the present study did not confirm the hypotheses, showing either no main effects of epistemic beliefs (eye-tracking data, Hypothesis 1a; verbal utterances, Hypothesis 3a) or even reversed effects (selection data, Hypothesis 2a; search outcome, Hypothesis 4a). However, as expected, differences between search interfaces were found for eye-tracking data (Hypothesis 1b), search result selection (Hypothesis 2b and 2c), and verbal utterances (Hypothesis 3b), but not for participants' search outcome (Hypothesis 4b). Furthermore, the assumption that the advantage of the tabular interface over the list interface would be larger for students with *high* beliefs that the Web contains correct knowledge was confirmed for the search outcome (Hypothesis 4c), and partly for verbal utterances (Hypothesis 3c), but not for eye-tracking data or search result selection (Hypothesis 1c and 2d). The following sections will discuss in detail the results of this study with regard to the effects of epistemic beliefs (7.4.1) and of the search interface (7.4.2) on participants' source evaluations on SERPs.

7.4.1 Relationships between Internet-specific epistemic beliefs and the evaluation of information quality

With regard to university freshmen's epistemic beliefs about whether or not the Web contains correct knowledge, unexpected patterns of results were obtained in the present study. Contrary to Hypotheses 1a and 3a and, thus, unlike the findings from Study 2, the results of the present study showed no significant relationships of epistemic beliefs with the total fixation time on search results or with the number of quality-related verbal utterances expressed on the SERPs. Furthermore, exploratory analyses revealed that the more students believed that the Web contains correct knowledge, the longer they fixated on search results that they did not select from the SERPs during their Web search. Contrary to the expectations, this suggests that beliefs that the Web contains correct knowledge were related to more thoughtful decisions to not select a search result in the present study. Moreover, unlike Study 2 that showed no relationship between epistemic beliefs and the selection of search results, in the present study the more students believed that the Web contains correct knowledge, the more objective Web pages they accessed and the more they tended to select an objective information source as the first search result.

To sum up, these findings indicate that, contrary to the results of Study 2, in the present study students who believed that the Web contains correct knowledge

showed a more appropriate Web search behavior when solving the given medical problem than students who had more doubts with regard to this issue.

A potential reason for the contradicting findings obtained in Study 2 and Study 3 of this dissertation might be that the items of the Internet-Specific Epistemological Questionnaire (ISEQ, Strømsø & Bråten, 2010) used to assess participants' epistemic beliefs in the two studies referred to different kinds of knowledge on the Web. Whereas the items used in the present study addressed Web-based knowledge in general (see Appendix H), those used in Study 2 of this dissertation or in the studies by Bråten and colleagues (Bråten et al., 2005; Strømsø & Bråten, 2010) addressed Web-based knowledge about study-related contents (see Appendix E). These different contexts seemed to influence study participants' interpretation of as well as their average agreement to the ISEQ items.

In Study 2 participants' average agreement with the ISEQ items was below the midpoint (i.e., 3) for all items (see Appendix E). This indicates that participants in general were rather skeptical about the quality of study-related knowledge available on the Web. The lowest ratings were obtained for the item "I am most confident that I have understood something for my classes when I have used the Internet as a resource" ($M = 1.41$, $SD = 0.67$). Considering that university students predominantly acquire scientific, study-related knowledge through quality-controlled resources such as lectures, textbooks, and scientific journals rather than from the Web (although some of these materials can be downloaded from the Web) epistemic doubts with regard to the quality of study-related knowledge available on the Web as compared to respective knowledge provided by other resources seem appropriate. Therefore, students who are more doubtful that the Web contains correct knowledge about study-related contents can be assumed to have a higher awareness of the challenges involved in finding high-quality study-related knowledge on the Web.

In contrast, in the present study, with the ISEQ items addressing Web-based knowledge in general, participants on average seemed to have a more positive attitude towards the Web as a knowledge resource, with average ratings above the midpoint (i.e., 3) obtained for five out of eight items (see Appendix H). The highest ratings were obtained for the items "The Internet contains accurate knowledge" ($M = 3.86$, $SD = 0.85$) and "The Internet can provide me with most of the knowledge I need to succeed in my daily life ($M = 4.05$, $SD = 0.98$). To answer these statements,

participants in the present study can be assumed to have thought about situations when using the Web to inform themselves about a certain topic (e.g., sports, politics, health, finance, news). Thus, their ratings are likely to be based on their previous experiences with the Web with regard to whether they found high-quality information or not. It is conceivable that students who have developed skills to find such high-quality information on the Web and to differentiate it from low-quality information might be those who highly agree with the ISEQ items. In that context, it seems plausible that students who believe that the Web (among other types of information) contains correct knowledge in the present study were those who accessed more objective Web pages. Further evidence for this assumption is given by a recent study by Kammerer, Amann, and Gerjets (2011) that examined non-academic adults (age range: 25-60 years) also searching the Web about therapies for Bechterew's disease (cf. Study 2 and 3 of the present dissertation). Results of this study indicated that the more participants' believed that the Web contains correct knowledge (as measured by the same items as used in Study 3), the more time they spent on objective Web pages and the less time they spent on subjective Web pages.

7.4.2 Effects of the search interface on the evaluation of information quality

In contrast to the unexpected results with regard to epistemic beliefs, the findings from the present study with regard to the search interface indicate that, in line with Hypotheses 1a and 2a, a tabular interface with labeled columns, in which search results are grouped according to different types of information sources, provides stronger affordances for source evaluations than a standard list interface, in that it supports the selection of objective (i.e., rather neutral, high quality) information and helps low-knowledge searchers to ignore potentially biased or one-sided commercial information. In particular, users of the tabular interface paid less attention to commercial search results on the SERPs (Hypothesis 1b) and accessed more objective Web pages and less commercial ones (Hypothesis 2b) than in the list interface. Contrary to Hypothesis 1a objective search results were not fixated longer in the tabular interface than in the list interface, but in line with the assumption that a tabular interface guides searchers with low prior domain knowledge in their Web search, users of the tabular interface more often selected an objective source as the first search result than users of the list interface (Hypothesis 2c). Furthermore, in line

with Hypothesis 3a, users of the tabular interface verbally referred more to the *general quality* of search results as users of the list interface. However, with regard to the number of more specific credibility-related utterances no differences were found between search interfaces. The effects of the search interface on verbal reflections related to the type of source or the search result position were moderated by participants' epistemic beliefs (see below). Furthermore, the expected positive effect of the tabular interface on users' search outcome with regard to the medical problem (Hypothesis 4a) was confirmed only for students with high beliefs that the Web contains correct knowledge, but not for students with doubts about that issue. Specifically, students with high beliefs that the Web contains correct knowledge included more arguments from objective Web pages in favor or against a therapy in their argumentative summaries when having used a tabular interface than when having used a list interface. In contrast, the tabular interface as compared to the list interface did not improve the search outcome of students with doubts that the Web contains correct knowledge. To conclude, the results indicate a combined effect of the tabular interface and high beliefs that the Web contains correct knowledge, rather than a compensation effect as was assumed in Hypothesis 4b.

Besides, the assumption that the increased attention to and selection of high-quality search results and the reduced attention to and selection of low-quality search results with the tabular interface would be larger for students with high beliefs that the Web contains correct knowledge was not confirmed (Hypothesis 1c and Hypothesis 2d). However, as expected, when using the tabular interface as compared to the list interface students with high beliefs that the Web contains correct knowledge referred less to the *position of the search results* (Hypothesis 3c). Furthermore, when using the tabular interface these students selected less search results than when using the list interface, which might reflect a more focused information selection. In contrast, students with doubts about the Web containing correct knowledge in both interfaces rarely referred to the position of the search results in the SERPs, but verbally reflected more on the *type of source* as well as tended to select more search results when using the tabular interface than when using the list interface. These findings might indicate that the students who had doubts that the Web contains correct knowledge also have doubted the classification provided by the tabular interface.

7.4.3 Conclusions

In summary, across different types of processing data as well as for the search outcome the results of the present study provide evidence that the search interface (i.e., the interface design of SERPs) and Internet-specific epistemic beliefs, or a combination of both, respectively, influence university students' source evaluations during Web search on a complex and unknown medical issue. Consistent with predictions, the presentation of Web search results by means of a tabular interface with additional source cues seems to support students with low prior domain knowledge in their search for and use of high-quality information. Yet, the tabular interface did not stimulate participants to verbally reflect more on specific credibility-related utterances and even though in the present study the number of participants who referred to source information in their argumentative summaries was slightly higher than in Study 1 and Study 2, such references were still quite rare and were not increased by the use of the tabular interface. Nonetheless, within the limits of their cognitive prerequisites, students in the present study seemed to evaluate the quality of Web information, as in both the list interface and the grid interface during their Web search they accessed more Web pages providing objective information than Web pages providing subjective or commercial information.

However, contrary to the expectations, the belief that the Web contains correct knowledge in the present study seemed to be beneficial rather than detrimental for selecting and processing high-quality objective information in order to solve the given medical problem. In particular, when using the tabular interface as compared to the list interface students with high beliefs that the Web contains correct knowledge produced better search outcomes in terms of the number of objective arguments included in their summaries. Thus, rather than possessing the inappropriate belief that *most* knowledge on the Web is correct, students who believe that the Web contains correct knowledge might be better characterized by the conviction that objective information *can* be found on the Web. In that context, it seems plausible that searchers with high beliefs that the Web contains correct knowledge might be eager to "separate the wheat from the chaff" by focusing on high-quality information and to ignore the rest, when solving the given medical problem on the Web. However, due to the ambiguity of the ISEQ items with regard to whether they mean that a) *all* or at least *most* knowledge on the Web is correct or that b) at least instances of correct knowledge *can* be found on the Web (cf. the

discussion in Chapter 5.1.3) it is not certain that participants in the present study have interpreted the items according to the second option. Moreover, one could question whether purely focusing on objective information sources really should be seen as the most appropriate search strategy to cope with controversially discussed topics. It could be argued that as long as Web users are aware of the fact that some information on the Web, for example, statements that come from subjective or commercial sources, might be unrepresentative or biased, taking this information into account can be quite valuable to acquire a comprehensive overview on the topic at hand. This implies that future research needs to be more differentiated with regard to the detailed circumstances under which a given epistemic belief about a certain type of Web-based knowledge of a person with a certain background is adaptive or not (cf. Hofer & Sinatra, 2010; Strømsø & Bråten, 2010). Additionally, the abovementioned ambiguity with regard to the scale "certainty and source of knowledge" of the Internet-Specific Epistemological Questionnaire (ISEQ, Strømsø & Bråten, 2010) should be clarified in future research. This could be done by asking participants how they interpret the items or by reformulating the items to make them less ambiguous.

Finally, some limitations of the present study should be addressed. First of all, participants in the present study were all university freshmen, constituting a homogeneous sample with Web search experience and skills higher than average. Second, the tabular interface used in the current study was an experimental mock-up. Although Finn and Kushmerick (2006) successfully classified a large set of online news sources in objective and subjective documents by using machine learning techniques, a reliable automatic classifications of search results according to these source categories might yet be difficult to generate computationally for the open Web. Moreover, differentiating Web pages with regard to the "objectivity" dimension of information quality, is only one possible way of categorization, which furthermore might have created an oversimplification of some users' concept of information sources on the Web. As has been outlined in Chapter 2.2 information quality is a multidimensional construct. Thus, information sources could alternatively or additionally be classified according to other dimensions such as authority or expertise, currency, or intended audience, just to mention a few. Besides, it has to be taken into account that the results of the present study concerning epistemic beliefs are only correlative in nature not warranting conclusions about causality.

8 General discussion

The World Wide Web has become a major resource for laypersons to search for medical and health information (cf. Fox, 2006; Morahan-Martin, 2004). However, as anyone can publish virtually anything on the WWW, the quality of medical and health information on the Web varies considerably, with many Web sites containing one-sided, biased, outdated, or even false information (Bates, Romina, Ahmed, & Hopson, 2006; Eysenbach, Powell, Kuss, & Sa, 2002). Thus, it has become more and more important to critically evaluate the quality of Web-based information sources and the information therein, in order to "separate the wheat from the chaff", that is, to filter out relevant and reliable information during Web search. Otherwise, searchers might potentially acquire doubtful or even false information.

Therefore, the purpose of the present dissertation was to investigate to what extent laypersons engage in source evaluations when searching the Web to find information about a complex medical or health-related issue (i.e., an ill-structured information problem). More precisely, this dissertation aimed at identifying factors that potentially facilitate the critical evaluation of information quality during Web search. Based on a conceptual framework proposed by Lazonder and Rouet (2008) three different types of variables that may influence the evaluation of information quality during Web search were examined, namely (1) *contextual variables*, (2) *individual variables*, and (3) *resource variables*. Specifically, the three empirical studies of this dissertation investigated the following factors: a) evaluation instructions given before the task as a contextual variable, b) searchers' prior domain knowledge and c) their Internet-specific epistemic beliefs as two individual variables, and d) the search interface a Web user has at his or her disposal to evaluate and select potentially relevant Web pages as a resource variable.

Section 8.1 explains briefly the rationale for selecting these factors and summarizes the general findings from the three studies of this dissertation. Then, in section 8.2 methodological considerations and the generalizability of the study results will be discussed. Finally, theoretical and practical implications of the findings from this dissertation will be presented in sections 8.3 and 8.4. Suggestions for future research will be provided in the context of sections 8.2, 8.3, and 8.4.

8.1   Summary of findings

As outlined in Chapter 3 of this dissertation, previous research on evaluation processes during Web search provides no conclusive evidence whether or under which preconditions laypersons spontaneously engage in evaluations of information quality when searching for science-related issues on the Web. Findings from several information science studies (Crystal & Greenberg, 2006; Law et al., 2006; Rieh, 2002; Tombros et al., 2005) indicate that Web searchers engage in evaluations of information quality to a substantial extent. However, the present dissertation proposed that these findings might be ascribed to explicit evaluation instructions (e.g., to mention during Web search the evaluation criteria one applies to evaluate search results and Web pages) given to the participants, who furthermore were likely to having possessed a certain amount of prior domain knowledge on the search topic.

Consistent with these assumptions, the central finding of the first study was that the combination of explicit evaluation instructions and a certain amount of prior domain knowledge resulted in increased evaluations of information quality during Web search on a complex health-related issue (i.e., about the effectiveness of low fat and low carb diets). University students with higher prior knowledge on diets and nutrition who received explicit evaluation instructions verbally reflected more on the credibility[4] and structure of information, tended to select more search results from the SERPs, and directed more gazes to source information displayed on the Web pages. Furthermore, the higher students' prior domain knowledge the more thoroughly they inspected the search results provided by the search engine and the more search results with lower rankings they selected from the SERPs. These effects, however, were independent of the evaluation instructions.

In contrast, students with no or little prior domain knowledge, irrespective of the instructions given, only rarely evaluated the quality of information and to a great extent simply relied on the ranking provided by the search engine predominantly selecting higher-ranked search results from the SERPs. This bears a severe risk for low-knowledge searchers to acquire doubtful or even false information.

---

[4] The definition of credibility utterances in Study 1 also comprised utterances related to the general quality and the type of source, which were treated as separate criteria in Studies 2 and 3.

Therefore, the purpose of the second and third study of the present dissertation was to examine other, more domain-independent factors that facilitate source evaluations of Web searchers with low prior domain knowledge. Two promising candidates identified were searchers' personal epistemology (as an individual variable) and the affordances of the search interface searchers have available to access information sources (as a resource variable). In particular, in the second and third study the focus was on the role of these two factors in low-knowledge searchers' source evaluations on SERPs, that is, on their decisions with regard to which Web pages to access for further inspection during Web search on a controversial medical issue. It was assumed that both a certain level of epistemic doubts with regard to the quality of information on the Web and a search interface that activates these doubts might be important preconditions for making evaluations of information quality during Web search on complex, unknown topics.

The underlying assumption of the present dissertation was that a search interface that presents search results in a vertical rank-ordered list and that displays only sparse and non-salient quality-related source information on the SERPs might not provide sufficient affordances to activate such doubts in low-knowledge searchers. Therefore, Study 2 compared a Google-like list interface to a grid interface that refrained from making the ranking of the search results the most salient feature by presenting the search results in a three-by-three grid. Furthermore, Study 3 compared a list interface to a tabular interface that aimed at both reducing the salience of the ranking and simultaneously increasing the salience of quality-related source information in the SERPs. In the tabular interface search results were grouped in three columns, labeled as "objective information", "subjective information", and "commercial information".

Results of the second study showed that the presentation of Web search results by means of a grid interface seems to support a more free exploration of the search results and to refrain searchers from focusing on only the first few results presented by the search engine. However, verbal data indicated that searchers' engagement in profound source evaluations regarding the type or credibility of information sources was indicated to also depend on their age. A potential reason for this effect might be that more profound evaluations of information quality require a certain amount of source knowledge which might grow with science-related search experience that students typically gain during their academic careers.

Furthermore, with regard to searchers' personal epistemology, the findings of the second study showed that participants with doubts that the Web contains correct study-related knowledge. independent of the search interface, visually inspected the search results more thoroughly and verbally reflected more on the type of source than students who were less skeptical about this issue. To conclude, these findings indicate that – at least to a certain extent – also in a standard list interface low-knowledge searchers' epistemic doubts with regard to the quality of Web information were activated. In addition, the grid interface seemed to compensate for a low epistemic awareness of the varying information quality, such that participants who were less doubtful that the Web contains correct study-related knowledge showed better search outcomes with the grid interface than with the list interface.

Yet, neither the use of the grid interface, nor epistemic doubts with regard to the quality of Web information led searchers to select the least trustworthy search results less often or to verbally reflect more on the credibility of information sources. These results indicate that low-knowledge searchers engaged in evaluations of information quality only to a moderate extent when evaluating search results, which might be due to the fact that the search results descriptions in Study 2 provided only sparse and non-salient quality-related information. In contrast, in Study 3 SERPs were augmented by quality-related information.

With regard to the effects of the search interface, the results of the third study indicate that the presentation of Web search results in a tabular interface with additional source cues supports low-knowledge searchers in their search for and use of high-quality Web information. In particular, in the tabular interface participants paid less attention to commercial search results on the SERPs and accessed more objective Web pages and less commercial ones than in the list interface.

However, with regard to searchers' epistemology, contrary to the second study the findings of the third study indicate that participants who believed that the Web contains correct knowledge showed a more appropriate Web search behavior when solving the given medical problem than those who had doubts about this issue. The higher participants' beliefs that the Web contains correct knowledge, the more objective Web pages they accessed and the longer they fixated on search results that they did not select from the SERPs during their Web search, indicating careful selection decisions to filter out high-quality information. Further, participants with

high beliefs that the Web contains correct knowledge achieved better search outcomes in the tabular interface than in the list interface, as measured by the number of arguments presented by objective Web pages that students pointed out in their summaries. As discussed in Chapter 7.4.1, the contradicting findings between the two studies with regard to epistemic beliefs might be explained by the fact that whereas Study 2 measured the extent to which participants believed that the Web contains correct knowledge about study-related contents, Study 3 assessed the extent to which participants believed that the Web contains correct knowledge in general. It is plausible that in the third study those participants who have developed skills to find high-quality information on the Web and to differentiate it from low-quality information were those being convinced that correct knowledge can be found on the Web.

Finally, it should be noted that, as in Study 2, in the third study neither the use of the tabular interface, nor epistemic beliefs that the Web contains correct knowledge led searchers to verbally reflect more on the credibility of information sources. This indicates that such profound evaluation processes or at least their verbalization might exceed the cognitive resources that low knowledge searchers have available to engage in source evaluations. Nonetheless, within the limits of users' individual cognitive and metacognitive prerequisites, a search interface that presents additional quality-related source information seems to lead to navigational decisions that are based to a substantial degree on evaluations of information quality.

To summarize, the three studies of the present dissertation provide evidence that all three types of variables, as classified in the conceptual framework by Lazonder and Rouet (2008), influence the evaluation of information quality during Web search: explicit evaluation instructions as a contextual variable, prior domain knowledge, Internet-specific epistemic beliefs, as well as age as individual variables, and the search interface users have at their disposal as a resource variable. However, it should be acknowledged that the present research comes with certain limitations that will be addressed in the next section.

8.2 Methodological considerations and generalizability of findings

The following sections address methodological considerations with regard to the thinking-aloud methods used in the three studies (section 8.2.1) and with regard to the measures used to assess the individual variables (section 8.2.2). Furthermore, considerations on the generalizability of the study results are presented in section 8.2.3.

8.2.1 Thinking-aloud methods: Concurrent thinking aloud versus cued retrospective reporting

To get insights into searchers' cognitive processes, in the present research two methods of thinking aloud have been used, namely concurrent thinking aloud (Study 1) and cued retrospective reporting (Study 2 and 3). In the following, the characteristics of both methods as they have been used in the studies of this dissertation as well as the assets and drawbacks of both methods will be outlined. It should be noted that evaluation instructions (used as experimental condition in Study 1) will not be addressed in the following discussion, because these instructions were not assumed to reflect spontaneous evaluation processes.

In Study 1 participants were asked to think aloud concurrently during their Web search. Specifically, in the condition with neutral thinking-aloud instructions in line with the standards described by Ericsson and Simon (1993) participants were asked to verbalize everything that came to their mind during task processing. Requiring participants to verbalize their thoughts concurrently to their search process, however, might have interfered with their cognitive processes and might have reactively influenced participants' search process itself. For instance, it is conceivable that even neutral thinking-aloud instructions (i.e., to verbalize everything that comes to one's mind) might have stimulated participants with higher prior knowledge to process the materials more elaborately. Alternatively, the same instructions might have additionally increased low-knowledge participants' cognitive burden during task processing (cf. Hertzum et al., 2009; Van Gog et al., 2008).

To avoid interferences of participants' verbalizations with their cognitive processes during Web search, in Study 2 and 3 of this dissertation, therefore, verbal data were obtained subsequent to the Web search task through the method of cued retrospective

reporting based on a replay of eye-movement recordings (cf. Hansen, 1991; Van Gog, Paas, Van Merriënboer, & Witte, 2005). Specifically, participants were presented a screen recording (at 50% speed) superimposed with their eye movements and mouse operations that were recorded during Web search and were asked each time when the screen recordings showed them visiting the Google SERPs to report what they were thinking during task processing. Thus, whereas concurrent thinking-aloud protocols directly reflect participants' thought processes during the Web search process, cued retrospective thinking-aloud protocols can only reflect the memory traces of the process. Furthermore, cued retrospective reporting bears the risk that participants' verbalizations are not only based on their memory, but also include post hoc explanations or fabrications (cf. Van Gog, Paas, Van Merriënboer, & Witte, 2005). To conclude, whereas verbal data obtained through the method of concurrent thinking-aloud (as used in the first study) can be assumed to reflect participants' spontaneous thought processes, but might alter their task processing, verbal data obtained through the method of cued retrospective reporting (as used in the second and third study) cannot influence the search process, but instead might include fabrications. These considerations should be taken into account when deciding on which of the two thinking-aloud methods to use in future studies. If it is of particular importance to examine users' natural, spontaneous Web search behavior, cued retrospective reporting might be the method of choice. In contrast, if it is of particular importance to obtain verbal data that reflect spontaneous on-line thought processes, concurrent thinking-aloud might be the preferred method. Prompting participants to explain their Web search and evaluation strategies, however, should be avoided in any case when studying search and evaluation processes, as this can activate processes that would not occur spontaneously (cf. Study 1).

8.2.2 Measures of prior domain knowledge and of epistemic beliefs

Prior domain knowledge and Internet-specific epistemic beliefs in the present research were assessed by means of rating scales comprising nine or eight items, respectively. The scale measuring prior domain knowledge achieved high internal consistency (9 items, see Appendix B; Cronbach's α = .86). However, it should be acknowledged that the three items rated highest by the participants, namely "It is important for me to eat healthy", "I'm interested in the issue of diets and nutrition", and "Diets and nutrition are an important issue with regard to my own health

condition" rather measure topic interest than prior domain knowledge (for the full list of items see Appendix x). Therefore, it is likely that participants with a higher prior domain knowledge in Study 1 also possessed a higher interest in the topic of low fat and low carb diets, which might have resulted in a higher motivation to perform well in the search task. Therefore, further research is needed to differentiate between effects of prior knowledge and interest on the topic on searchers' engagements in source evaluations during Web search. Related to that it should be acknowledged that motivational factors such as personal involvement (e.g., a medical patient recently diagnosed with a certain disease), which can be assumed to have a strong impact on searchers' engagement in evaluations of information quality (cf. Metzger, 2007) were not addressed in the present dissertation

With regard to the assessment of participants' epistemic beliefs about the extent to which they believed that the Web contains correct knowledge (as measured by the "certainty and source of knowledge" scale of the ISEQ by Strømsø and Bråten, 2010) as discussed in Chapter 7.4.1 the items used in Study 2 and Study 3 measured slightly different epistemic beliefs. Whereas the items used in Study 2 addressed Web-based knowledge about study-related contents (see Appendix E), those used in Study 3 addressed knowledge on the Web in general (see Appendix H). With respect to Web-based knowledge about study-related contents, it seems plausible that students who have doubts (i.e., low beliefs) that the Web contains correct knowledge about study-related contents have a higher awareness of the challenges involved in finding high-quality study-related knowledge on the Web. In contrast, with respect to Web-based knowledge in general, it is conceivable that students who have developed skills to find high-quality information on the Web and to differentiate it from low-quality information might be convinced that correct knowledge can be found on the Web (i.e., high beliefs). However, whereas both interpretations focus on high or low agreement with the statements, Hofer and Sinatra (2010) question whether the endpoints of Likert-type rating scales used to measure epistemic beliefs indeed should be interpreted as most or least appropriate beliefs. Instead, both a very high and a very low agreement with a statement such as "The Web contains accurate knowledge." might be rather absolutist. In contrast, the midpoint of the rating scale might measure the conviction that the Web contains both accurate and inaccurate knowledge, thus potentially representing a more appropriate perspective than both of the endpoints. However, the problem is that we cannot know how a participant

interprets the items (cf. Hofer & Sinatra, 2010). Yet, one possibility to shed light on participants' interpretation might be to not only provide items such as "The Web contains accurate knowledge", but also reversed items such as "The Web does not contain accurate knowledge" and even more important items that reflect a more relativistic view such as "The Web contains both accurate and inaccurate knowledge".

However, closely related to this issue of interpretation is the discussion of the domain-generality or domain-specificity of epistemic beliefs. Even though the present dissertation investigated Internet-specific epistemic beliefs instead of beliefs about the nature and knowledge in general or the nature and knowledge with regard to a specific domain or topic (e.g., medicine; weight loss methods), two questions remain unclear: First, whether in Study 2 participants' field of study had an impact on their degree of agreement to the statements that the Web contains correct knowledge about study-related contents and second, what kind of knowledge participants' had in mind when indicating their level of agreement to the statements that the Web contains correct knowledge. Thus, for future research that assesses Internet-specific epistemic beliefs with regard to study-related contents it might be advisable to involve participants from only one field of study (cf. Bråten et al., 2005; Strømsø & Bråten, 2010). When assessing epistemic beliefs with regard to Web-based knowledge in general, in the end of the questionnaire an open question might be added that requires participants to sketch what kind of knowledge they had in mind when rating the items.

Apart from considerations about possible improvements of questionnaires to assess epistemic beliefs, it should be noted that some researchers generally criticize the use of questionnaire-based measures to assess epistemic beliefs (e.g., diSessa, Elby, & Hammer, 2003; Louca, Elby, Hammer, & Kagey, 2004). These researchers claim that self-report questionnaires only measure epistemic beliefs in a decontextualized way, rather than in a given context. As epistemic beliefs can be activated in a certain context, but not in others, they postulate to assess them "in action", that is, as epistemic reflections or epistemic strategies brought to bear during task processing (cf. Hofer, 2004; Mason et al., 2010a; Mason, Boldrin, & Ariasi, 2010b; Mason, Ariasi, & Boldrin, 2011). However, in the present dissertation epistemic beliefs are regarded as separate from such ongoing epistemic reflections typically obtained from thinking aloud protocols. Whereas questionnaires can be assumed to target epistemic

179

beliefs as "mental states", thinking-aloud data collected during task processing might rather reflect "mental acts" (cf. Bråten & Strømsø, 2010). As stated above, thinking-aloud data in the present dissertation in contrast served as indicators for processes addressing the evaluation of information quality.

Finally, it has to be taken into account that the results of the present research concerning epistemic beliefs are only correlative in nature not warranting conclusions about causality. In this research epistemic beliefs were assessed one week after participants had conducted the Web search to avoid their Web search performance being affected by thoughts provoked by the questionnaire. Hence, there is a chance that participants' engagement in the Web search experiment might have influenced their epistemic beliefs. Indications for this assumption is given by recent findings from Kienhues, Stadtler, and Bromme (2011) who investigated the influence of the engagement in a medical Web search (about the topic of cholesterol) on university students' medicine-related beliefs. In their study, in the week before participants conducted the Web search and directly after the Web search participants' domain-specific epistemic beliefs about medicine were assessed. Results showed that after the Web search, participants viewed knowledge in medicine to be more imprecise or unstructured as well as more open or incomplete than before the Web search. In contrast, with regard to Internet-specific epistemic beliefs, a recent study by Kammerer, Amann, and Gerjets (2011) showed no influences of the engagement in a Web search experiment on participants' epistemic beliefs. In this study non-academic adults' Internet-specific epistemic beliefs were assessed one week before and one week after conducting a Web search about therapies for Bechterew's disease. Although subsequently to the Web search task participants even received a training on how to evaluate the quality of Web information, participants' epistemic beliefs about whether or not the Web contains correct knowledge (as measured by the same items as used in Study 3) did not significantly differ before and after the intervention. Accordingly, it seems rather unlikely that in the studies of the present dissertation participating in the Web search experiment changed participants' epistemic beliefs. Whereas Kienhues et al. (2011) administered the questionnaire immediately after the Web search, in the present research as well as in the work by Kammerer, Amann, and Gerjets (2011) it was administered one week after the experiment. Moreover, because in the present research experimental groups did not differ with regard to

epistemic beliefs, specific influences of the grid interface or the tabular interface on participants' epistemic beliefs can be excluded as well.

Besides, it should be noted that whereas the present research examined which role prior domain knowledge (Study 1) or epistemic beliefs (Study 2 and 3) played in searchers' source evaluations during Web search, the interrelations between the two individual variables were not investigated. This issue should be approached in future research.

8.2.3 Generalizability of findings

When drawing conclusions from the results of the present studies, the particular study conditions must necessarily considered. First of all, participants in the three studies of the present dissertation were all university students, constituting a homogeneous sample with computer- and Web search experience and skills higher than average. Second, for sake of experimental control some constraints were introduced in the studies. Participants conducted their Web search in a lab setting on an artificially designed search task with a predefined search time of 20 minutes (Study 1) or eight minutes (Study 2 and Study 3) and a finite set of search results comprising only 30 search results (Study 1) or 18 search results (Study 2 and Study 3), respectively. Furthermore, the experimental setting did not allow participants to enter their own search terms into the search engine or to take notes during their Web search. These study conditions might limit the generalizability of the results to a broader range of users (e.g., users with low Web search skills or without an academic background, older Web users, or school kids) as well as to other contexts such as more natural search situations with real information needs, without time constraints, and with the open Web at searchers' disposal. Furthermore, it is an open question whether the results found for topics like weight loss methods and therapies for Bechterew's disease are generalizable to other medical and health-related topics or to other controversially discussed science-related (e.g., climate change) or political issues (e.g., the withdrawal from nuclear energy). Finally, to test the effectiveness and user acceptance of alternative search interfaces such as those used in the present dissertation, long-term assessments in ecologically-valid settings with different types of users would be required.

Notwithstanding these limitations, the present dissertation provides novel insights into university students' evaluation processes during Web search on medical and health-related issues and the role of evaluation instructions, user characteristics, and the search interface in their engagement in evaluating the quality of Web information. The following section will elaborate in greater detail on the theoretical and practical implications of the present research.

8.3 Theoretical implications

The findings from the present research indicate that, when certain preconditions are given with regard to task instructions, users' individual prerequisites, or affordances of the search interface, laypersons' Web search on complex medical and health-related issues is guided to a substantial extent by evaluations of information quality. On the contrary, information foraging theory (Pirolli, 2007, Pirolli & Card, 1999; for details see Chapter 3.1), which is one of the most influential theories about the cognitive processes involved in hyperlink selection (e.g., the selection of search results) and navigating between several Web pages, is limited to evaluations of topical relevance. According to this theory, judgments of whether to select a search result or not for further processing and decisions when to leave a Web page are based on the notion of information scent, that is, the perceived topical relevance of screen objects to a user's current information need. Furthermore, satisficing strategies are assumed to play an important role in hyperlink selection. This implies that users do not evaluate the information scent of all search results available, but evaluate search results only until one is encountered that is "good enough", thus favoring the top positions. The evaluation of information quality, however, is ignored completely in this theory. This might be due to the fact that information foraging theory is rooted in research on simple fact-finding tasks or on tasks for which rather homogeneous sets of preselected, quality-controlled contents are provided. For these tasks information quality might not be an important issue.

However, the findings from the present dissertation indicate that when searchers deal with complex information problems for which information sources of highly variable quality can be found on the Web, search result selection and the processing of Web pages is also guided by aspects beyond topical relevance and link position. To conclude, for successfully modeling the selection of search results from SERPs in

Web search scenarios that address complex, ill-structured problems like those investigated in the present dissertation, information foraging theory might need to be extended by the following factors (see also Gerjets & Kammerer, 2010): First, source cues in search results or in Web pages that point to the quality or type of information (e.g., a ".org" domain name, or terms like "journal", "forum", or "shop") should be taken into account in addition to information scent cues. Furthermore, user characteristics, task instructions, as well as affordances of the search interface should be considered as important factors that influence users' attention to and interpretation of these source cues.

A theoretical framework that considers such factors is the documents model framework (see Chapter 3.2) that was proposed by Perfetti and colleagues (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006) and that was recently extended, for instance, by Rouet and Britt (2011) with their MD-TRACE model (see Chapter 3.2.1) or by Bråten, Britt, et al. (2011) with their integrated model of epistemic beliefs and documents model representation (see Chapter 5.1.2). The documents model framework predicts that competent readers who study multiple, diverse documents to learn about a complex topic at hand, engage in source evaluations by attending to and evaluating specific source characteristics of the documents (e.g., expertise or motives of a document's author), which allows them to interpret a document' content in the light of its source characteristics. However, up to now, empirical research that examined the concrete cognitive processes involved in source evaluations in the context of multiple-documents reading tasks is still rare. Most studies examined source evaluations only after participants had read the provided documents entirely, by asking participants to write essays on the topic at hand (e.g., Britt & Aglinskas, 2002), to judge the trustworthiness of the documents by means of rating scales (e.g., Bråten, Strømsø, & Salmerón, 2011; Rouet et al., 1996), or to select their favorite documents (e.g., Salmerón, Kammerer, et al., 2010). Moreover, in this type of research participants were usually provided with a relatively small set of preselected documents. In contrast, during Web search on complex issues, users themselves are responsible for selecting a manageable subset of the potentially most useful information sources for further exploration (cf. Braasch et al., 2009). Therefore, when applying the documents model framework on the Web context, source evaluations have not only to be considered during the evaluation of Web

pages, but also in an earlier stage of Web search, namely during the evaluation of search results (cf. Rouet & Britt, 2011).

Findings from the present dissertation confirm this assumption. Different types of processing data, such as eye tracking, log files, and verbal protocols, indicate that, given certain preconditions (see above), during Web search laypersons not only engage in source evaluations on Web pages, but also one step before, when deciding which search results to select from the SERPs for further inspections. Furthermore, the present research supports the integrated model of epistemic beliefs and documents model representation proposed by Bråten, Britt, et al. (2011), providing new evidence regarding the role of epistemic beliefs in the processing of source information.

To summarize, the present research has tied first empirical links between information foraging theory (Pirolli, 2007, Pirolli & Card, 1999) from cognitive science and the documents model framework (Britt et al., 1999; Perfetti et al., 1999; Rouet, 2006) and its extensions (Bråten, Britt, et al., 2011; Rouet & Britt, 2011) from text comprehension research. Integrating these two complementary theories that both address the evaluation of information sources, seems to be a promising way to better explain and predict laypersons' information search and evaluation processes on the Web when searching for complex and potentially controversial science-related issues. Future research might thus focus specifically on identifying what kind of verbal or pictorial cues on SERPs and on Web pages (either source information such as author information or information about the type of Web page, or semantic information indicating inconsistencies) trigger at which time during task processing evaluations of information quality for different types of searchers. Furthermore, with regard to user characteristics such as prior domain knowledge and epistemic beliefs further research is needed to provide deeper insights into *why* searchers with low prior domain knowledge or with inappropriate epistemic beliefs show deficient source evaluations during Web search. Moreover, whereas the present research only examined Web materials that varied with regard to information quality, but that were all of high topical relevance, future research is needed to examine the complex interplay between evaluations of topical relevance and of information quality. For example, it is an open question whether Web searchers in a first step evaluate the topical relevance of a piece of information and in a second step – in case that topical

relevance is given – the quality of the information or whether the reversed sequence is true.

8.4 Practical implications

From a practical point of view, the findings of the first study of this dissertation suggest that simple instructions that remind searchers to evaluate the information encountered during Web search have the potential to stimulate searchers to engage in source evaluations during Web search, given that they possess a certain amount of prior knowledge on the search topic at hand. Furthermore, the findings of the present research with regard to epistemic beliefs suggest, that even if prior domain knowledge is low, a high epistemic awareness about the diversity of information sources and the resulting necessity to critically evaluate the quality of information, facilitates source evaluations during Web search. Therefore, it seems crucial to create public awareness of the varying information quality on the Web, for example, through articles in magazines or newspapers that address this issue, or through trainings that inform school kids, university students, and working adults on how to evaluate information quality on the Web (e.g., Gerjets & Hellenthal-Schorr, 2008; Walraven et al., 2010; Wiley et al., 2009). In addition, the findings from the present dissertation suggest that the development of alternative search interfaces that provide salient quality-related source cues in the SERPs and/or display search results in a format different from a list is a promising way to facilitate laypersons' source evaluations during Web search on a complex issue. Furthermore, the issue of information quality seems also to be of current importance in the search engine industry. For instance, in an interview with Nature magazine in January 2010 about the future of Web search, Peter Norvig, the director of Google research underlined the importance of search engines implementing a measure of quality that is not only based on popularity. He pointed out that developing and improving search engine algorithms that determine both topical relevance (related to the user's query) and quality (related to factors independent of the query, such as the accuracy of information, or the trustworthiness of the authors) are a key challenge for the next decade. Furthermore, when looking at the current search interface of Google (as of March 2011) it becomes apparent that Google has already started to augment search results with quality-related information. For example, as shown in Figure 40 for search results linked to forum pages the number of posts, the number of authors, as

well as the date of the last post are displayed. Likewise, for search results about scientific articles the author name, publication year, how often the article was cited, and related articles are presented.
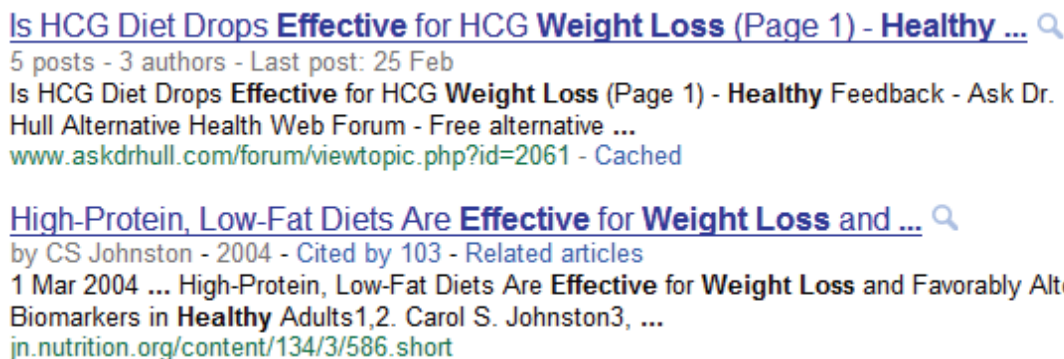
Figure 40. Google search results (as of March 2011) augmented by author, date, and citation information.

Moreover, by clicking on a magnifying glass presented next to each search result (see Figure 41), a small-scaled preview of the Web page is now made available in Google that may support the decision on whether to access a Web page or not. Finally, on the left side of the search interface, a fold-out menu with several filtering options is available (see also Figure 41), that allows searchers to only obtain search results linked to forum discussions ("Discussions) or news items about the topic ("News").
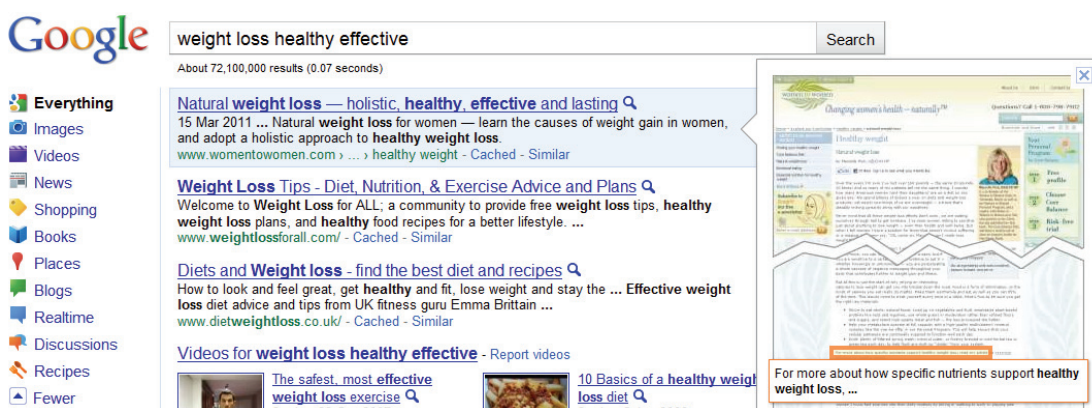
Figure 41. The Google SERP interface (as of March 2011) with a fold-out menu with several filtering options (presented to the left of the search results) and magnifying glasses to display small-scaled previews of the Web pages (presented to the right of the search results).

Future research in this area is needed to examine how these novel quality-related source cues or filters provided in the Google SERPs influence the search and evaluation behavior of Google users depending on their individual prerequisites and search goals.

To conclude, the present dissertation addresses a topic of great current interest for laypeople, educators, and search engine developers. Future research in this area might further contribute to a comprehensive understanding of how to help laypersons to "separate the wheat from the chaff", that is, to find high-quality information and ignore or devalue the rest when using the Web to find information about complex and controversially discussed issues.

References

Aiken, L. S., & West, S. G. (1991). *Multiple Regression: Testing and interpreting interactions.* Thousand Oaks: Sage.

Alexander, J. E., & Tate, M. A. (1999). *Web wisdom: How to evaluate and create information quality on the Web*. Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Anderson, J. R. (1983). A spreading activation theory of memory. *Journal of verbal learning and verbal behaviour, 22,* 261-295.

Aula, A., & Russell, (2008, June). *Complex and exploratory Web search.* Paper presented at the NSF IS3 Information Seeking Support Systems Workshop, Chapel Hill, NC, USA. Retrieved from http://www.ils.unc.edu/ISSS/papers/papers/aula.pdf

Barry, C. L., & Schamber, L. (1998). Users' criteria for relevance evaluation: A cross-situational comparison. *Information Processing & Management, 34*, 219-236.

Bateman, J. (1998). Changes in relevance criteria: A longitudinal study. In C. M. Preston (Ed.), *Proceedings of the American Society for Information Science* (ASIS 1998) (pp. 23–32). Medford, NJ: Information Today, Inc.

Bates, B. R., Romina, S., Ahmed, R., & Hopson, D. (2006). The effect of source credibility on consumers' perceptions of the quality of health information on the Internet. *Medical Informatics and the Internet in Medicine, 31,* 45-52.

Bendixen, L. D., & Hartley, K. (2003). Successful learning with hypermedia: The role of epistemological beliefs and metacognitive awareness. *Journal of Educational Computing Research, 28,* 15-30.

Borlund, P. (2003). The concept of relevance in IR. *Journal of the American Society for Information Science and Technology, 54*, 913-925.

Braasch, J., Lawless, K., Goldman, S., Manning, F., Gomez, K. & MacLeod, S. (2009). Evaluating search results: An empirical analysis of middle school students' use of source attributes to select useful sources. *Journal of Educational Computing Research*, *41,* 63-82.

Braasch, J., Rouet, J.-F., Britt, A., Knutsen, D., Le Bigot, L., & Vilbert, N. (2010, August). *Readers' construction of document models from news stories: Evidence from eye movements.* Paper presented at the Meeting of the EARLI Special Interest Group Text and Graphics Comprehension, Tübingen, Germany.

Brand-Gruwel, S., Van Meeuwen, L., & Van Gog, T. (2008). The use of evaluation criteria when searching the WWW: An eye-tracking study. In A. Maes & S. Ainsworth (Eds.), *Proceedings EARLI Special Interest Group Text and Graphics Comprehension 'Exploiting the opportunities: Learning with textual, graphical, and multimodal representations'* (pp. 34-37). Tilburg, The Netherlands: Tilburg University.

Brand-Gruwel, S., Wopereis, I., & Walraven, A. (2009). A descriptive model of information problem solving while using internet. *Computers & Education*, *53*, 1207-1217.

Bråten, I., Britt, M. A., Strømsø, H. I., & Rouet, J.-F. (2011). The role of epistemic beliefs in the comprehension of multiple expository texts: Towards an integrated model. *Educational Psychologist, 46,* 48-70.

Bråten, I., Gil, L., Strømsø, H. I., Vidal-Abarca, E. (2009). Personal epistemology across cultures: Exploring Norwegian and Spanish university students' epistemic beliefs about climate change. *Social Psychology of Education*, *12*, 529-560.

Bråten, I., & Strømsø, H. I. (2006a). Epistemological beliefs, interest, and gender as predictors of Internet-based learning activities. *Computers in Human Behavior, 22*, 1027-1042 .

Bråten, I., & Strømsø, H. I. (2006b). Effects of personal epistemology on the understanding of multiple texts. *Reading Psychology, 27*, 457-484.

Bråten, I., & Strømsø, H. I. (2010a). When law students read multiple documents about global warming: Examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instructional Science, 38*, 635-657.

Bråten, I., & Strømsø, H. I. (2010b). Effects of task instruction and personal epistemology on the understanding of multiple texts about climate change. *Discourse Processes, 47,* 1-31.

Bråten, I., Strømsø, H .I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly, 44*, 6-28.

Bråten, I., Strømsø, H. I., & Salmerón, L. (2011). Trust and mistrust when students read multiple information sources about climate change. *Learning and Instruction, 21*, 180-192.

Bråten, I., Strømsø, H. I., & Samuelstuen, M. S. (2005). The relationship between Internet-specific epistemological beliefs and learning within Internet technologies. *Journal of Educational Computing Research, 33*, 141-171.

Bråten, I., Strømsø, H.I., & Samuelstuen, M. S. (2008). Are sophisticated students always better? The role of topic-specific personal epistemology in the understanding of multiple expository texts. *Contemporary Educational Psychology, 33*, 814-840.

Britt, M. A., & Aglinskas, C. (2002). Improving students' ability to identify and use source information. *Cognition and Instruction, 20*, 485-522.

Britt, M. A., Perfetti, C. A., Sandak, R., & Rouet, J.-F. (1999). Content integration and source separation in learning from multiple texts. In S.R. Goldman, A.C. Graesser, & P. van den Broek (Eds.), *Narrative, comprehension, causality, and coherence: Essays in honor of Tom Trabasso* (pp. 209-233). Mahwah, NJ: Erlbaum.

Britt, M. A., & Rouet, J.-F. (2011). *Studying multiple documents: Cognitive processes and instructional implications.* Manuscript submitted for publication.

Broder, A. (2002). A taxonomy of Web search. *SIGIR Forum, 36*, 3-10.

Bromme, R., Kienhues, D., & Stahl, E. (2008). Knowledge and epistemological beliefs: An intimate but complicate relationship. In M. S. Khine (Ed.), *Knowing, knowledge and beliefs* (pp. 423-441). New York: Springer.

Brown, M. E. (1991). A general model of information-seeking behavior. In J.-M. Griffiths (Ed.), *Proceedings of the American Society for Information Science* (ASIS 1991) (pp. 9-14). Medford, NJ: Learned Information.

Brumby, D. P., & Howes, A. (2008). Strategies for guiding interactive search: An empirical investigation into the consequences of label relevance for assessment and selection. *Human-Computer Interaction, 23*, 1-46.

Buehl, M.M., & Alexander, P.A. (2001). Beliefs about academic knowledge. *Educational Psychology Review, 13,* 385-418.

Cho, J., & Roy, S. (2004). Impact of search engines on page popularity. In S. Feldman, & M. Uretsky (Eds.), *Proceedings of the 13th International Conference on World Wide Web* (WWW '04) (pp. 20-29). New York, NY: ACM Press.

Crystal, A., & Greenberg, J. (2006). Relevance criteria identified by health information users during Web searches. *Journal of the American Society for Information Science and Technology, 57*, 1368-1382.

Cutrell, E., & Guan, Z. (2007). What are you looking for? An eye-tracking study of information usage in Web search. In M. B. Rosson & D. Gilmore (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07) (pp. 407-416). New York, NY: ACM Press.

Danielson, D.R. (2005). Web credibility. In C. Ghaoui (Ed.), *Encyclopedia of Human-Computer Interaction* (pp. 713-721). Hershey, PA: Idea Group.

De Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2010). Attention guidance in learning from complex animation: seeing is understanding? *Learning and Instruction, 20*, 111-122.

Denzin, N. K. (1970). The research act in sociology. Chicago: Aldine.

Dillon, A. (2002).Writing as design: Hypermedia and the shape of information space. In R. Bromme & E. Stahl (Eds.), *Writing hypertext and learning: Conceptual and empirical approaches* (pp. 63-72). New York: Pergamon.

diSessa, A. A., Elby, A., & Hammer, D. (2003). J's epistemological stance and strategies. In G. M. Sinatra & P. R. Pintrich (Eds.), *Intentional conceptual change* (pp. 237–290). Mahwah, NJ: Lawrence Erlbaum Associates.

Dumais, S. T., Cutrell, E., & Chen, H. (2001). Optimizing search by showing results in context. In J. Jacko & A. Sears (Eds.), *Proceedings of the SIGCHI*

Conference on Human Factors in Computing Systems (CHI '01) (pp. 277-284). New York, NY: ACM Press.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data.* Cambridge, MA: MIT Press.

Erkens, G. (2005). *Multiple episode protocol analysis (MEPA)*. Version 4.10. The Netherlands: Utrecht University.

Eysenbach, G. (2008). Credibility of health information and digital media: New perspectives and implications for youth. In M. J. Metzger & A. J. Flanagin (Eds.), *Digital Media, Youth, and Credibility* (pp.123-154). The John D. and Catherine T. MacArthur Foundation.Cambridge, MA: The MIT Press.

Eysenbach, G., Powell, J., Kuss, O., & 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.

Fallows, D. (2008). Search engine use. *Pew Internet & American Life Project*. Washington, DC. Retrieved from http://www.pewinternet.org/Reports/2008/Search-Engine-Use.aspx

Fogg, B. J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P., & Treinen, M. (2001). What Makes A Web Site Credible? A Report on a Large Quantitative Study. In M. Beaudouin-Lafon & R. J. K. Jacob, *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '01) (pp. 62-68). New York, NY: ACM Press.

Fogg, B. J., Soohoo, C., Danielson, D. R., Marable, L., Stanford, J., & Tauber, E. R. (2003). How do users evaluate the credibility of Web sites? A study with over 2,500 participants. *In Proceedings of the 2003 Conference on Designing for User Experiences* (DUX '03) (pp. 1-15). New York, NY: ACM Press.

Fox, S. (2006). Online health search 2006. *Pew Internet & American Life Project*. Washington, DC. Retrieved from http://www.pewinternet.org/~/media//Files/Reports/2006/PIP_Online_Health_ 2006.pdf.pdf

Fox, S., & Jones, S. (2009). The social life of health information. *Pew Internet & American Life Project*. Washington, DC. Retrieved from http://www.pewinternet.org/Reports/2009/8-The-Social-Life-of-Health-Information.aspx

Fu, W-T. F., & Pirolli, P. (2007). SNIF-ACT: a cognitive model of user navigation on the World Wide Web. *Human Computer Interaction, 22,* 355-412.

Gerjets, P., & Hellenthal-Schorr, T. (2008). Competent information search in the World Wide Web: Development and evaluation of a web training for pupils. *Computers in Human Behavior, 24*, 693-715.

Gerjets, P., & Kammerer, Y. (2010). Topical relevance and information quality in cognitive models of Web search behavior: Introducing epistemic scent into information foraging theory. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 2278-2283). Austin, TX: Cognitive Science Society.

Gil, L., Bråten, I., Vidal-Abarca, E., & Strømsø, H.I. (2010). Summary versus argument tasks when working with multiple documents: Which is better for whom? *Contemporary Educational Psychology, 35, 157-173*.

Granka, L., Joachims, T., & Gay, G. (2004). Eye-tracking analysis of user behavior in www search. In M. Sanderson (Ed.), *Proceedings of the 27th Annual ACM SIGIR Conference on Research and Development in Information Retrieval* (SIGIR '04) (pp. 478-479). New York, NY: ACM Press.

Guan, Z., & Cutrell, E. (2007). *An eye tracking study of the effect of target rank on Web search.* In M. B. Rosson & D. Gilmore (Eds.), *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI '07) (pp. 417-420). New York, NY: ACM Press.

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

Harris, R. (2010). Evaluating Internet research sources. Retrieved from http://www.virtualsalt.com/evalu8it.htm

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.

Hirsh, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science, 50*, 1265-1283.

Hofer, B. K. (2004). Epistemological understanding as a metacognitive process: Thinking aloud during online searching. *Educational Psychologist, 39*, 43-55.

Hofer, B. K., & Pintrich, P. R. (1997). The development of epistemological theories: Beliefs about knowledge and knowing and their relation to learning. *Review of Educational Research, 67,* 88-140.

Hofer, B.K., & Sinatra, G.M. (2010). Epistemology, metacognition, and self-regulation: Musings on an emerging field. *Metacognition and Learning, 4,* 113-120.

Holsanova, J., Holmberg, N., & Holmqvist, K. (2008). Reading information graphics: The role of spatial contiguity and dual attentional guidance. *Applied Cognitive Psychology, 22*, 1-12.

Hölscher, C., & Strube, G. (2000). Web search behavior of Internet experts and newbies. *Computer Network, 33*, 337-346.

Horrigan, J. (2006). The internet as a resource for news and information about science. *Pew Internet & American Life Project.* Retrieved from http://www.pewinternet.org/Reports/2006/The-Internet-as-a-Resource-for-News-and-Information-about-Science.aspx

Hwang, K. O., Farheen, K., Johnson, C. W., Thomas, E. J., Barnes, A. S., & Bernstam, E. V. (2007). Quality of weight loss advice on internet forums. *American Journal of Medicine, 120*, 604-609. Retrieved from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1995087/?tool=pmcentrez

Ivory, M. Y., Yu, S., & Gronemyer, K. (2004). Search result exploration: A preliminary study of blind and sighted users' decision making and performance. In E. Dykstra-Erickson, & K. M. Tscheligi (Eds.), *CHI '04 Extended Abstracts on Human Factors in Computing Systems* (CHI '04) (pp. 1453-1456). New York, NY: ACM Press.

Jacob, R. J. K., & Karn, K. S. (2003). Eye tracking in Human-Computer Interaction and usability research: Ready to deliver the promises, In J. Hyönä, R. Radach,

& H. Deubel (Eds.), *The mind's eye: Cognitive and applied aspects of eye movement research* (pp. 573-605). Amsterdam: Elsevier.

Jarodzka, H., Scheiter, K., Gerjets, P., & Van Gog, T. (2010). In the eyes of the beholder: how experts and novices interpret dynamic stimuli. *Learning and Instruction, 20*, 146-154.

Joachims, T., Granka, L., Pan, B., Hembrooke, H., & Gay, G. (2005). *Accurately interpreting clickthrough data as implicit feedback.* In R. A. Baeza-Yates, & N. Ziviani (Eds.), *Proceedings of the 28th Annual international ACM SIGIR Conference on Research and Development in information Retrieval* (SIGIR '05) (pp. 154-161). New York, NY: ACM Press.

Jonassen, D. H. (1997). Instructional design models for well-structured and ill-structured problem-solving learning outcomes. *Educational Technology: Research and Development, 45*, 65-94.

Josephon & Holmes (2002). Visual attention to repeated internet images: testing the scanpath theory on the world wide web. In A. T. Duchowski (Ed), *Proceedings of the 2002 Symposium on Eye Tracking Research & Applications* (ETRA '02) (pp. 43-49). New York, NY: ACM Press.

Juvina, I., & Van Oostendorp, H. (2008). Modeling semantic and structural knowledge in Web navigation. *Discourse Processes, 45*, 346-364.

Kammerer, Y. (2010, May). *Further improvement of an offset correction method to overcome the inaccuracy of fixation data*. Poster presented at the Scandinavian Workshop on Applied Eye-Tracking (SWAET '10), Lund, Sweden.

Kammerer, Y., Amann, D., & Gerjets, P. (2011, July). *The role of internet-specific epistemic beliefs in non-academic adults' web search for medical information*. Paper to be presented at the 21th annual meeting of the Society for Text and Discourse, Poitiers, France.

Kammerer, Y., Wollny, E., Gerjets, P., & 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* (pp. 2158-2163). Austin, TX: Cognitive Science Society.

Kammerer, Y., Wollny, E., & Gerjets, P. (2011). How source cues in Web search results and domain-general epistemic beliefs affect Web search for health information. Manuscript in preparation.

Karanam, S., Oostendorp, H. van, Puerta Melguizo, M.C., & Indurkhya, B. (2009). Integrating graphical information into cognitive modelling of web-navigation. In N. A. Taatgen & H. van Rijn (Eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (pp. 2807-2812). Austin, TX: Cognitive Science Society.

Keane, M.T., O'Brien, M., & Smyth, B. (2008). Are people biased in their use of search engines? *Communications of the ACM, 51,* 49-52.

Kellar, M., Watters, C., & Shepherd, M. (2007). A field study characterizing Web-based information-seeking tasks. *Journal of the American Society for Information Science and Technology, 58,* 999-1018.

Kienhues, D., Stadtler, M. & Bromme, R. (2011). Dealing with conflicting or consistent medical information on the Web: When expert information breeds laypersons' doubts about experts. *Learning and Instruction, 21*, 193-204.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review, 95,* 163-182.

Kintsch, W. (1998). *Comprehension: A paradigm for cognition.* Cambridge, MA: Cambridge University Press.

Kitajima, M., Blackmon, M. H., & Polson, P. G. (2000). A comprehension-based model of Web navigation and its application to Web usability analysis. In S. Mc-Donald, Y. Waern, & G. Cockton (Eds.), *People and computers XIV— Usability or else!* (pp. 357-373). New York: Springer.

Kuhlthau, C. (1991). Inside the search process: Information seeking from the user's perspective. *Journal of the American Society for Information Science, 42,* 361-371.

Kuhlthau, C. (1993). *Seeking meaning. A process approach to library and information services*. Norwood, N.J.: Ablex.

Law, E. L-C., Klobucar, T., & Pipan, M. (2006). User effect in evaluating personalized information retrieval systems. In W. Nejdl & K. Tochtermann

(Eds), *Proceedings of the First European Conference on Technology-enhanced Learning* (EC-TEL 2006) (257-271). Berlin/Heidelberg: Springer.

Lazonder, A. W., & Rouet, J.-F. (2008). Information problem solving instruction: Some cognitive and metacognitive issues. *Computers in Human Behavior, 24*, 753-765.

Le Bigot, L., & Rouet, J-.F. (2007). The impact of presentation format, task assignment, and prior knowledge on students' comprehension of multiple online documents. *Journal of Literacy Research, 39*, 445-470.

Lewandowski, D. (2011). The influence of commercial intent of search results on their perceived relevance. In *Proceedings of the 2011 iConference* (iConference '11) (pp. 452-458). New York, NY: ACM Press.

Lorigo, L., Pan, B., Hembrooke, H., Joachims, T., Granka, L., & Gay, G. (2006). The influence of task and gender on search and evaluation behavior using Google. *Information Processing and Management, 42*, 1123-1131.

Louca, L., Elby, A., Hammer, D., & Kagey, T. (2004). Epistemological resources: Applying a new epistemological framework to science instruction. *Educational Psychologist, 39*, 57–68.

MaKinster, J. G., Beghetto, R. A., & Plucker, J. A. (2002). "Why can't I find Newton's Third Law? Case studies of students using of the Web as a science resource." *Journal of Science Education and Technology*, *11*, 155-172.

Mansell, P., & Read, J. (2009). Posttraumatic stress disorder, drug companies, and the Internet. *Journal of Trauma & Dissociation, 10*, 9-23.

Marchionini, G., & White, R. W. (2008). Find What You Need, Understand What You Find. *Journal of Human-Computer Interaction, 23*, 205–237.

Marchionini, G. (1989). Information-seeking strategies of novices using a full-text electronic encyclopedia. *Journal of the American Society for Information Science, 40*, 54–66.

Marchionini, G. (1992). Interfaces for end-user information seeking. *Journal of the American Society for Information Science, 43*, 156–163.

Marchionini, G. (1995). *Information seeking in electronic environments*. Cambridge: Cambridge University Press.

Martin, M. J. (2010). Reliability and type of consumer health documents on the world wide web: an annotation study. In *Proceedings of the NAACL HLT 2010 Second Louhi Workshop on Text and Data Mining of Health Documents* (Louhi '10) (pp. 38-45). Stroudsburg, PA: Association for Computational Linguistics.

Mason, L., & Ariasi, N. (2010). Critical thinking about biology during web page reading: Tracking students' evaluation of sources and information through eye fixations. In L. Verschaffel, E. De Corte, T. de Jong, & J. Elen (Eds.), *Use of external representations in reasoning and problem solving: Analysis and improvement* (pp. 55-73). London: Routledge.

Mason, L., Ariasi, N., & Boldrin, A. (2011). Epistemic beliefs in action: Spontaneous reflections about knowledge and knowing during online information searching and their influence on learning. *Learning and Instruction, 21,* 137-151.

Mason, L., Boldrin, A., & Ariasi, N. (2010a). Epistemic metacognition in context: Evaluating and learning online information. *Metacognition and Learning, 5,* 67-90.

Mason, L., Boldrin, A., & Ariasi, N. (2010b). Searching the Web to learn about a controversial topic: Are students epistemically active? *Instructional Science, 38*, 607-633.

Metzger, M. J. (2007). Making sense of credibility on the Web: Models for evaluating online information and recommendations for future research. *Journal of the American Society for Information Science and Technology, 58,* 2078–2091.

Miller, C. S., & Remington, R. W. (2004). Modeling information navigation: Implications for information architecture. *Human-Computer Interaction*, *19*, 225-271.

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

Muis, K.R., Bendixen, L.D., & Haerle, F.C. (2006). Domain-generality and domain-specificity in personal epistemology research: Philosophical and empirical reflections in the development of a theoretical framework. *Educational Psychology Review, 18*, 3-54.

Navarro-Prieto, R., Scaife, M., & Rogers, Y. (1999). Cognitive strategies in Web searching. In *Proceedings of the 5th Conference on Human Factors & the Web*. Retrieved from http://zing.ncsl.nist.gov/hfWeb/proceedings/navarro-prieto/index.html.

Nokes, J. D., Dole, J. A., & Hacker, D. J. (2007). Teaching high school students to use heuristics while reading historical texts. *Journal of Educational Psychology, 99,* 492-504.

O'Brien, M., & Keane, M.T. (2006). Modeling result-list searching in the World Wide Web: The role of relevance topologies and trust bias. In R. Sun, & N. Miyake (Eds.), *Proceedings of the 28th Annual Conference of the Cognitive Science Society* (pp. 1881-1886). Mahwah, NJ: Erlbaum.

Opfermann, M. (2008). There's more to it than instructional design - The role of individual learner characteristics for hypermedia learning. Berlin: Logos.

Opinion: 2020 visions (2010). *Nature, 463*, 26-32. doi:10.1038/463026a. Retrieved from http://www.nature.com/nature/journal/v463/n7277/full/463026a.html

Pan, B., Hembrooke, H., Joachims, T., Lorigo, L., Gay, G., & Granka, L. (2007). In Google we trust: Users' decisions on rank, position, and relevance. *Journal of Computer-Mediated Communication, 12*, 801–823.

Perfetti, C. A., Rouet, J.-F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. van Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading*. Mahwah, NJ: Erlbaum.

Pieschl, S., Stahl, E., & Bromme, R. (2008). Epistemological beliefs and self-regulated learning with hypertext. *Metacognition and Learning, 3*, 17-37.

Pirolli, P., Card, S. K. (1995). Information foraging in information access environments. In *Proceedings of the CHI '95 ACM Conference on Human Factors in Software* (pp. 51–58), New York: ACM Press.

Pirolli, P. (2007). *Information Foraging Theory. Adaptive interaction with information*. New York: Oxford University Press.

Pirolli, P., & Card, S. K. (1999). Information Foraging. *Psychological Review, 106*, 643-675.

Pirolli, P., & Fu, W-T. F. (2003). SNIF-ACT: a model of information foraging on the world wide Web. In P. Brusilovsky, A. Corbett, & F. de Rosis (Eds.), *9th International Conference on User Modeling* (pp. 45-54). Johnstown, PA: Springer.

Rayner, K. (1998). Eye movements in reading and information processing: 20 years of research. *Psychological Bulletin, 124*, 372-422.

Richter, T. (2007). Wie analysiert man Interaktionen von metrischen und kategorialen Prädiktoren? Nicht mit Median-Splits! [How to analyze interactions of metric and categorical predictors: Not with median splits!] *Zeitschrift für Medienpsychologie, 19*, 116-125.

Rieh, S. Y. (2002). Judgment of information quality and cognitive authority in the Web. *Journal of the American Society for Information Science and Technology, 53*, 145-161.

Rieh, S. Y. & Danielson, D. R. (2007). Credibility: A multidisciplinary framework. In B. Cronin (Ed.), *Annual Review of Information Science and Technology* (Vol. 41, pp. 307-364). Medford, NJ: Information Today.

Rieh, S. Y., & Hilligoss, B. (2008). College students' credibility judgments in the information-seeking process. In M. J. Metzger, & A. J. Flanagin (Eds.). *Digital Media, Youth, and Credibility. The John D. and Catherine T. MacArthur Foundation Series on Digital Media and Learning* (pp. 49–72). Cambridge, MA: The MIT Press.

Rose, D. E., & Levinson, D. (2004). Understanding user goals in Web search. In S. Feldman, & M. Uretsky (Eds.), *Proceedings of the 13th International Conference on World Wide Web* (WWW '04) (pp. 13-19). New York, NY: ACM Press.

Rouet, J.-F. (2006). *The skills of document use: From text comprehension to Web-based learning*. Mahwah, NJ: Erlbaum.

Rouet, J.-F., & Britt, M. A. (2011). Relevance processes in multiple document comprehension. In M.T. McCrudden, J. P. Magliano, & G. Schraw (Eds.), *Text relevance and learning from text* (pp. 19–52). Greenwich, CT: Information Age Publishing.

Rouet, J.-F., Britt, M.A., Caroux, L., Nivet, C., & Le Bigot, L. (2009). *The influence of story consistency and reading context on the construction of documents models from multiple sources.* Paper presented at the biennial conference of the European Association for Research on Learning and Instruction, Amsterdam, The Netherlands.

Rouet, J.-F., Britt, M.A., Mason, R.A., & Perfetti, C.A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88*, 478-493.

Rouet, J.-F., Favart, M., Britt, M. A., & Perfetti, C. A. (1997). Studying and using multiple documents in history: Effects of discipline expertise. *Cognition and Instruction, 15*, 85-106.

Salmerón, L., Gil, L., Bråten, I., & Strømsø, H. I. (2010). Comprehension effects of signalling relationships between documents in search engines. *Computers in Human Behavior, 26*, 419-426.

Salmerón, L., Kammerer, Y., Llorens, A., & Garcia-Carrión, P. (2010, August). *In Google we trust? Domain knowledge and students' Web page selections when searching for a controversial topic.* Paper presented at the Meeting of the EARLI SIG2 Text and Graphics Comprehension, Tübingen, Germany.

Sanchez, C. A., Wiley, J., & Goldman, S.R. (2006). Teaching students to evaluate source reliability during Internet research tasks. In S.A. Barab, K.E. Hay, & D.T. Hickey (Eds.), *Proceedings of the Seventh International Conference on the Learning Sciences* (pp. 662-666). Bloomington, IN: International Society of the Learning Sciences.

Sankoff & Kruskal (1983). *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison.* Addison-Wesley, Reading, MA.

Saracevic, T. (2007a). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science & Technology, 58*, 1915-1933.

Saracevic, T. (2007b). Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of

relevance. *Journal of the American Society for Information Science & Technology, 58*, 2126-2144.

Savolainen, R., & Kari, J. (2006). User-defined relevance criteria in Web searching. *Journal of Documentation, 62*, 685-707.

Schamber, L. (1994). Relevance and information behaviour. In M. E. Wiliams (Ed.), *Annual review of information science and technology* (ARIST 1994) (pp. 3-48). Medford, NJ: Learned Information.

Scheiter, K., & Van Gog, T. (2009). Using eye tracking in applied research to study and stimulate the processing of information from multi-representational sources. *Applied Cognitive Psychology, 23*, 1209-1214.

Schommer, M. (1990). Effects of beliefs about the nature of knowledge on comprehension. *Journal of Educational Psychology, 82*, 498-504.

Schraw, G., Dunkle, M. E., & Bendixen, L. D. (1995). Cognitive processes in well-defined and ill-defined problem solving. *Applied Cognitive Psychology, 9*, 523-538.

Shneiderman, B. (1997) Designing information-abundant Web sites: issues and recommendations. *International Journal of Human-Computer Studies, 47*, 5-29.

Shneiderman, B., Byrd, D., & Croft, W. B. (1997). Clarifying Search: A User-Interface Framework for Text Searches. *DL Magazine.* Retrieved from http://www.dlib.org/dlib/january97/retrieval/01shneiderman.html

Shneiderman, B., Byrd, D., & Croft, W. B. (1998). Sorting out searching: A user-interface framework for text searches. *Communications of the ACM, 41*, 95–98.

Simon, H. A. (1955). A behavioral model of rational choice. *Quarterly Journal of Economics, 69*, 99–118.

Smith, A. (2009). The internet's role in campaign 2008. *Pew Internet & American Life Project*. Washington, DC. Retrieved from http://www.pewinternet.org/Reports/2009/6--The-Internets-Role-in-Campaign-2008.aspx

Stadtler, M., & Bromme, R. (2007). Dealing with multiple documents on the WWW: The role of metacognition in the formation of documents models. *Computer-Supported Collaborative Learning, 2*, 191-210.

Stadtler, M., & Bromme, R. (2008). Effects of the metacognitive tool met.a.ware on the Web search of laypersons. *Computers in Human Behavior, 24*, 716–737.

Stanford, J., Tauber, E. R., Fogg, B. J., & Marable, L. (2002). Experts vs. online consumers: A comparative credibility study of health and finance Web sites. *Consumer WebWatch Research Report*. Retrieved from http://www.consumerwebwatch.org/dynamic/web-credibility-reports-experts-vs-online.cfm

Strømsø, H. I., & Bråten, I. (2010). The role of personal epistemology in the self-regulation of internet-based learning. *Metacognition and Learning, 5*, 91-111.

Strømsø, H. I., Bråten, I., & Britt, M. A. (2010). Reading multiple texts about climate change: The relationship between memory for sources and text comprehension. *Learning and Instruction, 20,* 192-204.

Strømsø, H.I., Bråten, I. & Samuelstuen, M.S. (2008). Dimensions of topic-specific epistemological beliefs as predictors of multiple text understanding. *Learning and Instruction, 18,* 513-527.

Sundar, S. S., Knobloch-Westerwick, S., & Hastall, M. R. (2007). News cues: Information scent and cognitive heuristics. *Journal of the American Society for Information Science and Technology, 58*, 366-378.

Sutcliffe, A. G., & Ennis, M. (1998). Towards a cognitive theory of information retrieval. *Interacting with Computers*, *10*, 321–351.

Suthers, D. & Hundhausen, C. D. (2003). An experimental study of the effects of representational guidance on collaborative learning processes. *The Journal of the Learning Sciences, 12*, 183-218.

Symonds, P. M. (1936). Influence of order of presentation of items in ranking. *Journal of Educational Psychology, 27*, 445-449.

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* (pp. 194-204). Amsterdam: IOS Press.

Tate, M .A. (2010). *Web wisdom: How to evaluate and create information quality on the Web* (2nd ed.). Boca Raton, FL: CRC Press.

Tombros, A., Ruthven, I., & Jose, J. M. (2005). How users assess Web pages for information-seeking. *Journal of the American Society for Information Science and Technology*, *56*, 327-344.

Tu, Y. W., Shih, M., & Tsai, C. C. (2008). Eighth graders' web searching strategies and outcomes: The role of task types, web experiences and epistemological beliefs. *Computers and Education*, *51*, 1142-1153.

Van Eimeren, B. & Frees, B. (2009). Der Internetnutzer 2009 – multimedial und total vernetzt? Media Perspektiven 07/2009. Retrieved from http://www.ard-zdf-onlinestudie.de/fileadmin/Online09/Eimeren1_7_09.pdf

Van Gog, T., Brand-Gruwel, S., Van Meeuwen, L., & Paas, F. (2008). Uncovering cognitive processes: Cued retrospective reporting based eye movement records. In *Proceedings of the International Conference of the Learning Sciences* (ICLS '08). Utrecht, NL, 416-417.

Van Gog, T., Paas, F., & Van Merriënboer, J. J. G. (2005). Uncovering expertise-related differences in troubleshooting performance: combining eye movement and concurrent verbal protocol data. *Applied Cognitive Psychology, 19*, 205-221.

Van Gog, T., Paas, F., Van Merriënboer, J. J. G., & 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., & Boshuizen, H. P. A. (2008). Information-problem solving: A review of problems students encounter and instructional solutions. *Computers in Human Behavior, 24*, 623-648.

Walraven, A., Brand-Gruwel, S., & Boshuizen, H. P. A. (2010). Fostering transfer of web searchers' evaluation skills: A field test of two transfer theories. *Computers in Human Behavior, 26,* 716-728.

Wang, P., & Soergel, D. (1998). A cognitive model of document use during a research project: Study I. Document selection. *Journal of the American Society for Information Science, 49,* 115–133.

Whitmire, E. (2004). The relationship between undergraduates' epistemological beliefs, reflective judgment, and their information-seeking behaviour. *Information Processing and Management, 40,* 97-111.

Wiley, J., Goldman, S., Graesser, A., Sanchez. C., Ash, I. & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal. 46,* 1060-1106.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation, 55*, 249-270.

Wineburg, S. S. (1991). Historical problem solving: A study of the cognitive processes used in the evaluation of documentary and pictorial evidence. *Journal of Educational Psychology*, *83*, 79.

Winter, S., Krämer, N. C., Schielke, K., & Appel, J. (2010). Information selection in the blogosphere. The effect of expertise, community rating, and age. In S. Ohlsson & R. Catrambone (Eds.), *Proceedings of the 32nd Annual Conference of the Cognitive Science Society* (pp. 802-807). Austin, TX: Cognitive Science Society.

Wopereis, I., Brand-Gruwel, S., & Vermetten, Y. (2008). The effect of embedded instruction on solving information problems. *Computers in Human Behavior, 24,* 738-752.

YouGovPsychonomics AG (2009). *Health Care Monitoring 2009*. Retrieved from http://www.presseportal.de/pm/25608/1500348/yougovpsychonomics_ag

Zusammenfassung

Das World Wide Web (WWW) ist in den letzten Jahren zu einer zentralen Informationsressource für medizinische und gesundheitsbezogene Themen geworden. Laut Umfragen der PEW Internet & American Life Project (Fox & Jones, 2009) und der Studie "Health Care Monitoring" (2009) des deutschen Marktforschungsinstituts YouGovPsychonomics AG suchten im Jahre 2009 83% der amerikanischen Internetnutzer und 79% der deutschen Internetnutzer im Web nach medizinischen und gesundheitsbezogenen Informationen. Viele von ihnen sind Laien, die sich beispielsweise über verschiedene Therapiemöglichkeiten zur Behandlung einer bestimmten Krankheit oder über Diät- oder Ernährungsfragen informieren (Fox, 2006; Morahan-Martin, 2004). Oftmals geht es also nicht um das Auffinden einfacher, eindeutiger Fakten, sondern um komplexe, kontrovers diskutierte Themen. Zudem können die Informationen, die der Internetnutzer während seiner Recherche im Web findet, maßgeblich seine Entscheidung, z.B. sich einer bestimmten Therapie zu unterziehen oder eine bestimmte Diätmethode anzuwenden, beeinflussen (Fox, 2006). Aufgrund der Tatsache, dass jeder nahezu jede denkbare Information im Web veröffentlichen kann, variiert die Qualität der medizinischen und gesundheitsbezogenen Informationen im Web jedoch stark: Neben seriösen, wissenschaftlich fundierten Webseiten gibt es auch eine Vielzahl an Webseiten, die unqualifizierte, einseitige oder falsche Informationen bereitstellen. Folglich ist es wichtig, dass Internetnutzer nicht nur die inhaltliche Relevanz von Informationen für ihr Recherchethema, sondern auch die Qualität der Informationen, auf die sie im Web stoßen, kritisch bewerten, um die "Spreu vom Weizen zu trennen".

Daher war das Ziel der vorliegenden Dissertation, zu untersuchen, unter welchen Voraussetzungen und in welchem Ausmaß Laien bei der Web-Recherche zu einem komplexen und kontrovers diskutierten medizinischen oder gesundheitsbezogenen Thema die vorgefundenen Web-Informationen kritisch hinsichtlich ihrer Qualität bewerten. Als potentielle Einflussfaktoren auf die Bewertung der Informationsqualität während der Webrecherche wurden basierend auf einem konzeptuellen Modell von Lazonder und Rouet (2008) drei verschiedene Arten von Variablen betrachtet: (a) Kontextvariablen, (b) individuelle Variablen und (c)

Ressourcenvariablen. Konkreter Gegenstand der vorliegenden Dissertation bildete die theoretische Analyse und empirische Untersuchung von Bewertungsinstruktionen als Kontextvariable, domänenspezifischem Vorwissen und epistemologischen Überzeugungen als individuelle Variablen und Affordanzen des Suchmaschinen-Interface als Ressourcenvariable. Bewertungsprozesse der Probanden (Studenten) bei der Webrecherche zu einem komplexen Thema (in Studie 1 Diätmethoden; in Studie 2 und 3 Therapiemöglichkeiten bei Morbus Bechterew) wurden mittels Blickbewegungs- und Logfile-Analysen, sowie verbalen Protokollen untersucht. Darüber hinaus wurde auch das Ergebnis der Informationsrecherche betrachtet.

Kapitel 2 zeigt die Wichtigkeit von Qualitätsbewertungen bei der Websuche zu medizinischen Themen auf und beschreibt Informationsqualität als ein multidimensionalen Konstrukt. Des Weiteren wird die Informationssuche im Web als eine Art Problemlöseprozess definiert (vgl. Marchionini, 1995; Brand-Gruwel et al., 2009). Anschließend wird anhand des IPS-I-Modells (*I*nformation *P*roblem *S*olving on the *I*nternet) nach Brand-Gruwel et al. (2009) beschrieben, welche kognitiven und metakognitiven Prozesse während der Informationssuche im Web ablaufen. Abschließend werden in Anlehnung an Rieh (2002) und Hilligoss und Rieh (2008) drei Phasen unterschieden, in denen Bewertungsprozesse während der Informationssuche stattfinden können, nämlich (1) die Bewertung von Suchergebnissen auf den SERPs (search engine results pages), (2) die Bewertung einer Webseite und (3) die Bewertung einer Sammlung von Webseiten.

In **Kapitel 3** werden zwei zentrale Theorien, welche die Bewertung von Informationsquellen adressieren, die Information Foraging Theorie (Pirolli, 2007; Pirolli & Card, 1999) aus der Kognitionswissenschaft sowie die Theorie der Repräsentation multipler Dokumente (Perfetti et al., 1999) aus der Textverstehensforschung, vorgestellt. Die Information Foraging Theorie bezieht sich vor allem auf die ersten beiden Bewertungsphasen, wohingegen die Theorie der Repräsentation multipler Dokumente insbesondere die zweite und dritte Bewertungsphase adressiert. Es wird erläutert, dass die Information Foraging Theorie die Auswahl von Suchergebnissen und die Rezeption von Webseiten unter dem Gesichtspunkt der thematischen Relevanz von Informationen erklärt (d.h. Bewertung der Informationsqualität bleibt unberücksichtigt). Die Theorie der Repräsentation multipler Dokumenten postuliert hingegen, dass kompetente Leser

207

Quellenbewertungen vornehmen (z.B. anhand des Autors, des Datums, des Dokumententyps, etc.), um die Informationsqualität eines Dokuments zu bewerten und entsprechend die Aussagekraft bestimmter Informationen zu bestimmen. Des Weiteren werden in **Kapitel 3** empirische Befunde aus den Informationswissenschaften dargestellt, die eine Reihe von qualitätsbezogenen Bewertungskriterien aufzeigen, die Webnutzer während der Informationssuche im Web anwenden. Sowohl theoretische Überlegungen als auch empirische Ergebnisse aus der Forschung zum Verstehen multipler Dokumente deuten allerdings darauf hin, dass dieses recht hohe Ausmaß an qualitätsbezogenen Bewertungskriterien auf ein gewisses Maß an Vorwissen (individuelle Variable) in Kombination mit expliziten Bewertungsinstruktionen (Kontextvariable) zurückzuführen ist.

Diese Annahme wurde in der ersten Studie der vorliegenden Dissertation (**Kapitel 4**) empirisch überprüft. Die Studie untersuchte den Einfluss von einer expliziten Bewertungsinstruktion (mit der Aufforderung während der Websuche die Bewertungskriterien zu nennen, die man verwendet, um Suchergebnisse und Webseiten zu bewerten) und von domänenspezifischem Vorwissen auf die Bewertungsprozesse bei der Webrecherche. Übereinstimmend mit den Vorhersagen zeigten die Ergebnisse der Studie, dass eine explizite Bewertungsinstruktion gegenüber einer neutralen Laut-Denken-Instruktion qualitätsbezogene Bewertungsprozesse von Studenten mit einem gewissen Maß an Vorwissen sowohl auf SERPs als auch auf Webseiten signifikant erhöhte. Studenten mit niedrigem Vorwissen bewerteten die Informationsqualität auf SERPs und auf Webseiten jedoch unabhängig von der Instruktion, die sie erhielten, nur sehr selten. Anstatt dessen tendierten sie dazu sich einfach auf das Ranking der Suchmaschine zu verlassen.

Aufgrund dieser Befunde war das Ziel des weiteren Teils der vorliegenden Arbeit zu untersuchen, ob es bestimmte domänenunabhängigere Voraussetzungen gibt, unter denen selbst Webnutzer mit niedrigem Vorwissen bezüglich des Recherchethemas die Informationen kritisch hinsichtlich ihrer Qualität bewerten (insbesondere auf den SERPs). **Kapitel 5** beschreibt theoretische Überlegungen sowie empirische Befunde, die darauf hindeuten, dass epistemologische Überzeugungen (individuelle Variable) und das Suchmaschinen-Interface (Ressourcenvariable) eine wichtige Rolle für die Bewertung der Informationsqualität bei der Webrecherche spielen könnten. **Kapitel 5.1** definiert epistemologische Überzeugungen, d.h. die persönliche Annahmen einer

Person über die Natur des Wissens und dessen Entstehung, als ein mehrdimensionales und mehrebiges Konstrukt. Des Weiteren wird eine Erweiterung der Theorie der Repräsentation multipler Dokumente vorgestellt (Bråten, Britt, et al., 2011), die Annahmen über die Rolle epistemologische Überzeugungen beim Aufbau mentaler Repräsentationen über multiple Dokumente macht und wie beispielsweise epistemologische Überzeugungen mit einer kritischen Bewertung von Quelleninformationen zusammenhängen. Bezüglich Internet-spezifischer epistemologischer Überzeugungen, d.h. Annahmen über die Natur des Wissens und dessen Entstehung im Web, wird in Anlehnung an Strømsø et al. (2005) angenommen, dass Webnutzern mit der Überzeugung, dass das Web eine reliable Informationsquelle ist, die korrektes Wissen enthält, die Gefahr unseriöse oder falsche Informationen zu rezipieren, nicht bewusst ist. Demnach sollten Nutzer bei der Webrecherche die Informationsqualität im Web umso kritischer bewerten, je mehr Zweifel sie bezüglich der Qualität von Web-Informationen haben. **Kapitel 5.2** zeigt allerdings zwei zentrale Eigenschaften von herkömmlichen Suchmaschinen-Interfaces auf, die dafür verantwortlich sein könnten, das mögliche Zweifel bezüglich der Qualität von Webinformationen während der Webrecherche nicht "aktiviert" werden, nämlich (1) die Präsentation der Suchergebnisse in einer vertikalen, nach inhaltlicher Relevanz geordneten Liste und (2) das Fehlen oder die Insalienz von qualitätsbezogenen Informationen in den Suchergebnissen.

Daher testete die zweite Studie der vorliegenden Dissertation (**Kapitel 6**) die Annahme, dass ein Suchmaschinen-Interface, das die Suchergebnisse nicht in einer vertikalen Liste sondern in einem Gitterformat (d.h. in einer 3x3-Matrix) präsentiert, Webnutzer mit niedrigem Vorwissen mehr zu eigenständigen Bewertungsprozessen bei der Auswahl von Suchergebnissen anregt, als ein herkömmliches Listeninterface (vgl. Google). Dies sollte insbesondere der Fall sein für Webnutzer, die Zweifel daran haben, dass das Web korrektes Wissen (über studiumsbezogene Inhalte) enthält. Um zu untersuchen, inwiefern die Probanden (alle hatten niedriges Vorwissen) während der Webrecherche bei der Auswahl der Suchergebnisse auf die Qualität der Informationen achten, wurde folgendes methodisches Paradigma gewählt: Die Reihenfolge der Suchergebnisse auf einer SERP wurde experimentell variiert, indem sie entweder in optimaler Reihenfolge (Suchergebnis mit höchster Vertrauenswürdigkeit zuerst) oder in umgekehrter Reihenfolge (Suchergebnis mit niedrigster Vertrauenswürdigkeit zuerst) präsentiert wurden. Ergebnisse der Studie

zeigten, dass sowohl das Gitter-Interface als auch epistemologische Zweifel, dass das Web korrektes, studiumsbezogenes Wissen enthält zu verstärkten qualitätsbezogenen Bewertungsprozessen führten, insbesondere wenn die obersten Suchergebnisse eher unseriös waren. In diesem Fall wurden beispielsweise mit dem Gitter-Interface mehr vertrauenswürdige Suchergebnisse ausgewählt als mit dem Listen-Interface. Zusätzlich scheint das Gitter-Interface eine unkritische Überzeugung gegenüber der Qualität von studiumsbezogenen Web-Informationen zu kompensieren, da Studenten mit solch einer Überzeugung mit dem Gitter-Interface ein besseres Ergebnis der Informationsrecherche erzielten als mit dem Listen-Interface. Allerdings konnten weder das Gitter-Interface noch Zweifel bezüglich der Qualität von Web-Informationen bewirken, dass die Studenten weniger unseriöse Suchergebnisse auswählten oder mehr über die Glaubwürdigkeit von Informationen reflektierten, was daran liegen könnte, dass die verwendeten Suchmaschinen-Interfaces beide nur wenig und insaliente qualitätsbezogene Informationen präsentierten.

Deshalb wurde in der dritten Studie (**Kapitel 7**) ein Listeninterface mit einem Tabellen-Interface verglichen, in dem die Suchergebnisse nach objektiven, subjektiven und kommerziellen Informationen geordnet waren. Die Ergebnisse der Studie zeigten, dass ein Tabellen-Interface die Studenten (alle hatten niedriges Vorwissen) in der Bewertung der Informationsqualität und somit bei der Suche nach und Auswahl von neutralen, qualitativ hochwertigen Informationen unterstützte. Bezüglich der epistemologischen Überzeugungen zeigten die Ergebnisse allerdings, dass die Überzeugung, dass das Web korrektes Wissen enthält, mit einer erhöhten Auswahl an Suchergebnissen von objektiven Quellen und einer intensiveren Betrachtung nicht-ausgewählter Suchergebnisse einherging (unabhängig vom Suchmaschinen-Interface), was auf sorgfältige Entscheidungsprozesse zur Auswahl qualitativ hochwertiger Informationen schließen lässt. Außerdem schienen insbesondere Studenten mit dieser epistemologischen Überzeugung von dem Tabellen-Interface zu profitierten, da sie mit dem Tabellen-Interface bessere Recherche-Ergebnisse erzielten als mit einem Listen-Interface. Diese zunächst widersprüchlich erscheinenden Ergebnisse aus Studie 2 und Studie 3 werden in **Kapitel 7** (und **Kapitel 8**) diskutiert und auf die unterschiedliche Spezifität der Items (bezogen auf studiumsbezogenes Wissen im Web vs. bezogen auf Wissen im Web allgemein), die zur Erfassung der Internet-spezifischen epistemologischen Überzeugungen in den beiden Studien verwendet wurden, zurückgeführt.

Abschließend erfolgt in **Kapitel 8** eine Zusammenfassung und Diskussion der Ergebnisse der drei Studien. Des Weiteren werden methodische Gesichtspunkte der drei Studien sowie die Generalisierbarkeit der Befunde, theoretische und praktische Implikationen der Erkenntnisse, und Vorschläge für weiterführende Forschung diskutiert. Als Fazit wird gezogen, dass alle drei Arten von Variablen nach dem konzeptuellen Modell von Lazonder und Rouet (2008), genauer gesagt, explizite Bewertungsinstruktionen, domänenspezifisches Vorwissen und Internet-spezifische epistemologische Überzeugungen, sowie Affordanzen des Suchmaschinen-Interface eine wichtige Rolle bei der Bewertung der Informationsqualität bei der Webrecherche spielen. Die Ergebnisse der Studie zeigen, dass unter bestimmten Voraussetzungen sowohl bei der Bewertung von Webseiten, aber auch ein Schritt zuvor, bei der Auswahl der Suchergebnisse auf den SERPs qualitätsbezogene Bewertungsprozesse ablaufen. Solche Voraussetzungen sind beispielsweise ein gewisses Maß an Vorwissen in Kombination mit der Aufforderung Suchergebnisse und Webseiten zu bewerten, oder bei niedrigem Vorwissen ein gewisses Maß an epistemologischer Bewusstheit über die Variabilität der Informationsqualität im Web, oder ein angemessenes Suchmaschinen-Interface, das Affordanzen zur Bewertung der Informationsqualität bietet.

Appendix A

Search results used in Study 1

Figure A1. Screenshot of the SERP with "low carb" as keywords.

Figure A2. Screenshot of the SERP with "low carb + low fat" as keywords.

Appendix B

Items of the prior knowledge scale about diets and nutrition

(with means and standard deviations for each item)

Instructions:

Please rate the following statements. It is your personal opinion that interests us. Thus, there are no right or wrong answers.

Statements	*M*	*SD*
It is important for me to eat healthy.	3.66	1.14
I know more about diets and nutrition than my family and friends.	2.62	0.90
I'm interested in the issue of diets and nutrition.	3.17	1.20
I have never heard about the low carb versus low fat controversy. (r)	2.86	1.62
Diets and nutrition are an important issue with regard to my own health condition.	3.48	1.02
I could spontaneously list a number of low carb diets.	1.38	0.68
I could spontaneously list a number of low fat diets.	1.34	0.72
I can describe precisely the concept of low carb diets.	1.69	0.93
I can describe precisely the concept of low carb diets.	1.76	1.06

Note. All statements were answered on a 5-point scale ranging from 1 (*I totally disagree*) to 5 (*I totally agree*). Statements marked with an "r" were re-coded.

Appendix C

Items of the "computer- and Web search experience and skills" scale

1. How do you rate your computer skills?

very low (1) – *very high* (5)

2. How do you rate your skills to search information on the Web?

very low (1) – *very high* (5)

3. How often do you use the Web to search for information?

very seldom or never (1) – *very often* (5)

Appendix D

Search results used in Study 2

Figure D1. Screenshot of the list interface (optimal trustworthiness order) for the query "bechterew's disease radon".
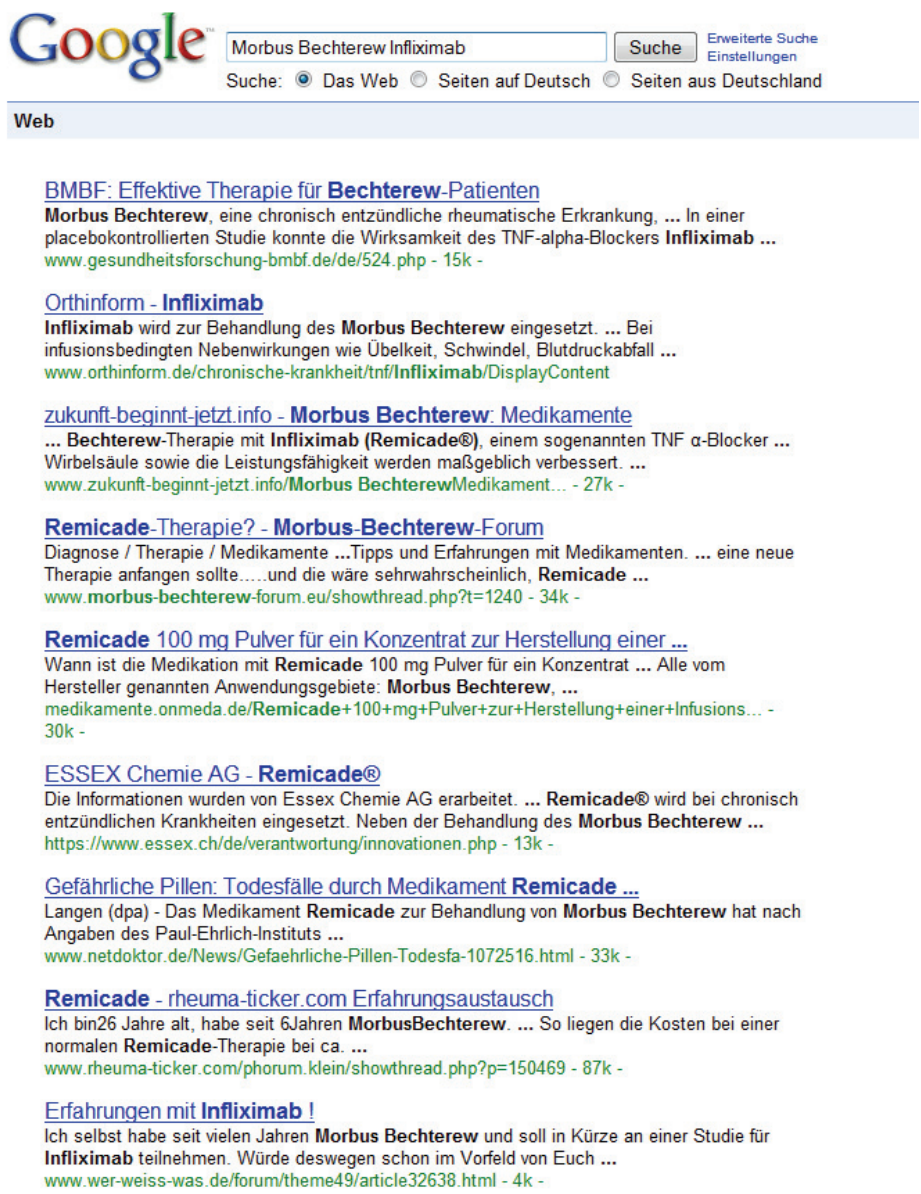
Figure D2. Screenshot of the list interface (optimal trustworthiness order) for the query "bechterew's disease infliximab".

Appendix E

Items of the Internet-specific epistemic beliefs scale about the Web containing correct knowledge about study-related contents

(with means and standard deviations for each item)

Instructions:

Please rate the following statements that concern the study-related knowledge that exists on the Internet and the Internet as a knowledge resource. It is your personal opinion that interests us. Thus, there are no right or wrong answers.

Statements	*M*	*SD*
The Internet contains accurate knowledge about the topics I study	2.91	1.08
The Internet can provide me with most of the knowledge I need to succeed in my courses.	2.29	1.03
The truth about almost every issue raised in my classes is located on the Internet.	2.19	1.00
On the Internet many different sources provide the correct answer to questions related to my course work.	2.77	0.99
I am most confident that I have understood something for my classes when I have used the Internet as a resource.	1.41	0.67
Most of what is true in my field of study is available on the Internet.	2.27	1.01
When I encounter difficult problems in my course work, I feel I am on safe ground if I find expert statements about them on the Internet.	2.77	1.12
The correct answer to questions in my course work exists on the Internet.	2.25	0.84

Note. All statements were answered on a 5-point scale ranging from 1 (*I totally disagree*) to 5 (*I totally agree*).

Appendix F

Items of the prior knowledge scale about Bechterew's disease or other rheumatic diseases and respective therapies

1. I'm familiar with rheumatic diseases.

2. I have never heard about the Bechterew's disease. (r)

3. I know the typical symptoms of Bechterew's disease.

4. I know therapies to treat Bechterew's disease.

5. I have never heard of the agent "infliximab". (r)

6. I have never heard about a radon therapy. (r)

7. I have never heard of the drug "remicade". (r)

8. I can explain precisely what a healing gallery (Heilstollen) is.

9. I can explain precisely what a tnf-alpha blocker is.

10. I can explain precisely what a balneotherapy is.

All statements were answered on a 5-point scale ranging from 1 (*I totally disagree*) to 5 (*I totally agree*). Statements marked with an "r" were re-coded.

Appendix G

Search results used in Study 3

Figure G1. Screenshot of the list interface (random order) for the query "bechterew's disease radon".
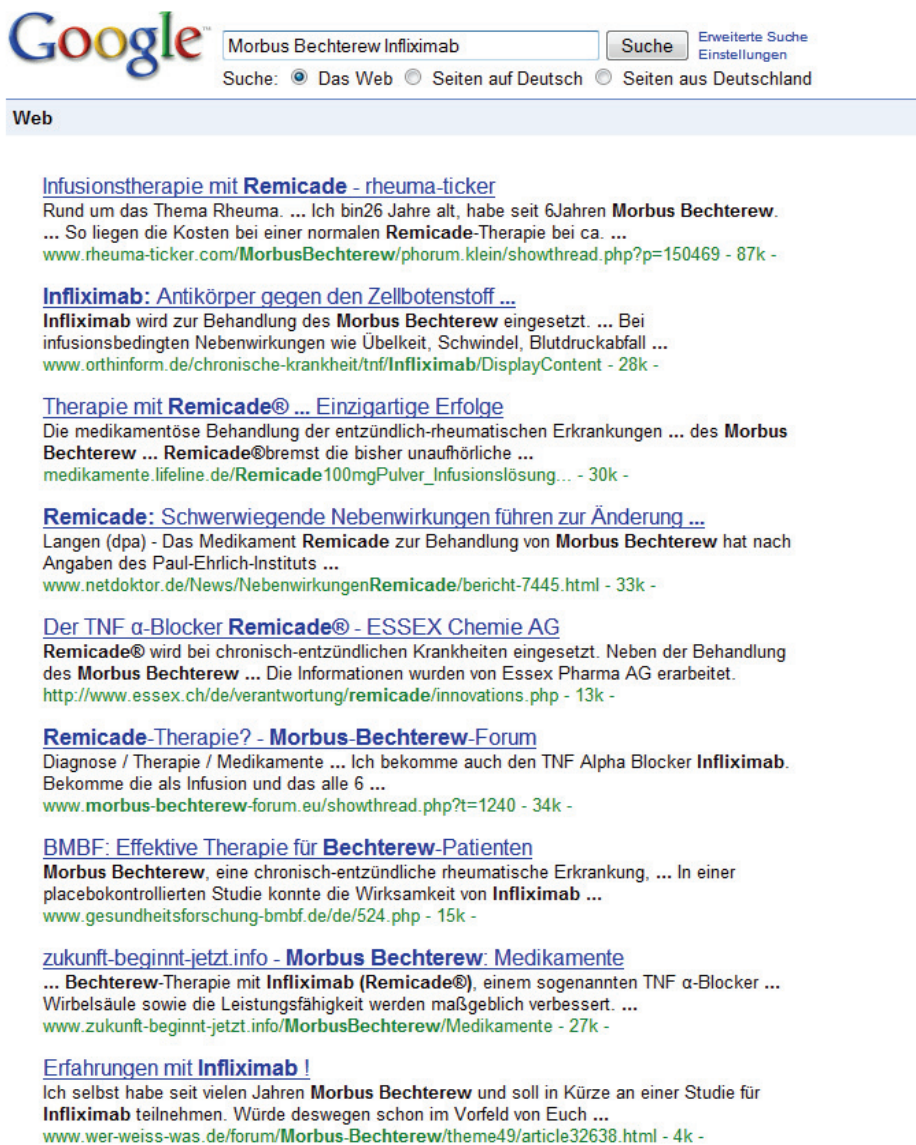
Figure G1.   Screenshot of the list interface (random order) for the query "bechterew's disease infliximab".

Appendix H

Items of the Internet-specific epistemic beliefs scale about the Web containing correct knowledge

(with means and standard deviations for each item)

Instructions:

Please rate the following statements that concern the knowledge that exists on the Internet and the Internet as a knowledge resource. It is your personal opinion that interests us. Thus, there are no right or wrong answers.

Statements	*M*	*SD*
The Internet contains accurate knowledge	3.86	0.85
The Internet can provide me with most of the knowledge I need to succeed in my daily life.	4.05	0.98
The truth about almost every conceivable topic is located on the Internet.	2.91	1.13
On the Internet many different sources provide the correct answer to questions I have about a given topic.	3.81	0.83
I am most confident that I have understood something about a given topic when I have used the Internet as a resource.	2.24	0.80
Most of what is true about a given topic is available on the Internet.	2.53	1.05
When I encounter difficult problems, I feel I am on safe ground if I find expert statements about them on the Internet.	3.26	1.05
The correct answer to my questions exists on the Internet.	3.40	1.06

Note. All statements were answered on a 5-point scale ranging from 1 (*I totally disagree*) to 5 (*I totally agree*).