Computer-Supported Collaborative Learning (2007) 2:191–210
DOI 10.1007/s11412-007-9015-3

Dealing with multiple documents on the WWW: The role of metacognition in the formation of documents models

Marc Stadtler · Rainer Bromme

Received: 8 September 2006 / Accepted: 10 July 2007 /
Published online: 7 September 2007

Abstract Drawing on the theory of documents representation (Perfetti et al., Toward a theory of documents representation. In: H. v. Oostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading*. Mahwah, NJ: Erlbaum, 1999), we argue that successfully dealing with multiple documents on the World Wide Web requires readers to form documents models; that is, to form a representation of contents *and* sources. We present a study in which we tested the assumption that the use of metacognitive strategies is crucial to the formation of documents models. A total of 100 participants with little medical knowledge were asked to conduct an Internet research on a medical topic. Participants were randomly assigned to four experimental groups that received different types of metacognitive prompts: participants either received evaluation prompts, monitoring prompts, both types of prompts, or no prompts. A control group took paper-and-pencil notes. Results showed that laypersons receiving evaluation prompts outperformed controls in terms of knowledge about sources and produced more arguments relating to the source of information when justifying credibility judgments. However, laypersons receiving evaluation prompts were not better able to indicate the source of information after Internet research than controls. In addition, laypersons receiving monitoring prompts acquired significantly more knowledge about facts, and performed slightly better on a comprehension test. It is concluded that the results underline the importance of metacognition in dealing with multiple documents.

Keywords Comprehension of multiple documents · Metacognition · Metacognitive tools · Internet research · Expert–layperson-communication

Introduction

With the rising dissemination of scientific information on the Internet, learning from the World Wide Web (WWW) has become a popular activity both in formal education as well

M. Stadtler (✉) · R. Bromme
Psychology Department, University of Muenster, Fliednerstraße 21, 48149 Muenster, Germany
e-mail: stadtlm@uni-muenster.de

R. Bromme
e-mail: bromme@uni-muenster.de

as outside of schools and academic contexts. On the web, learners have immediate access to a wealth of information comprised of differing standpoints and which—due to the speed of publishing—is often more up to date than the knowledge represented in books or scientific journals.

A widespread example of informal learning from the WWW is the research for medical information conducted by laypersons. Laypersons often access health information on the WWW to learn about a specific disease or different treatment alternatives, especially in the run-up to important health-related decisions. The information they retrieve may help them to make a knowledge-based decision—something that is commonly taken to be an important precondition for patient compliance (O'Connor 1995). The resulting learning situation differs from traditional learning settings in that laypersons certainly do not aim to become experts, yet need to develop a basic understanding of the relevant concepts (Bromme et al. 2005).

However, even when the information is available, laypersons may find it hard to deal with its complexity and heterogeneity. Relevant information is scattered across a multitude of different web sites (Bhavnani et al. 2003), making it necessary to integrate information; that is, to forge semantic links between information from different sources. This process may be hampered by a lack of textual cues, such as transitional statements clarifying the relation between different bits of information, which are usually provided by authors in single texts (Goldman and Rakestraw 2000).

As well as the contents, laypersons have to deal with the sources of information (Hofer 2004). Awareness about source information is particularly important when dealing with medical information on the WWW, because "gatekeepers of credibility," such as editors and publishers are missing (Britt and Aglinskas 2002). As a consequence, numerous studies have documented severe quality deficits in medical information provisions (see, for a review, Eysenbach et al. 2002).

To summarize, dealing with scientific health-related information on the WWW is an interesting and important example of learning from multiple documents, an issue that has mostly been analyzed up to now in the academic context of schools and universities and with reference to printed documents (e.g., Britt and Aglinskas 2002; Rouet et al. 1996; Wineburg 1991).

Theoretical background

Dealing with multiple documents: The theory of documents representation

Traditionally, research on how readers comprehend and represent written text has focused on the case of reading single texts (e.g., Kintsch 1998; Kintsch and van Dijk 1978). However, readers often need to deal with more than one text, for example, when learning about a controversial historical issue or a complex scientific field in which different views or different pieces of information have to be gathered from different documents.

Recognizing the need to adapt traditional propositional models of text representation to the situation of multiple documents, Perfetti et al. (1999) have developed the "theory of documents representation." Basically, this theory describes a text representation called "documents model" that the authors deem most appropriate for dealing with multiple documents. The documents model is made up of two interconnected, yet separately accessible mental representations of the documents a reader deals with, i.e., the situations model and the intertext model. In the situations model, readers represent contents from the documents they are dealing with in an integrated format. These may take the form of causal-temporal chains of arguments, as is illustrated in Fig. 1. In this fictitious example, a reader mentally represented the situation described in document 1 (http://www.food-industries.com), i.e., eating high-cholesterol food enhances one's blood-cholesterol level, which in turn enhances the risk of developing a coronary heart disease. This information is supplemented by a second document (http://www.dr-clark.org), from which the reader derived the information that the largest part of cholesterol is produced endogenously, i.e., in the human liver. She further represented from that second document that refraining from high-cholesterol food thus has only a marginal effect on the blood cholesterol level.

In the intertext model, both information about the sources of the documents and the relationship between documents is represented. Source information is stored in the form of document nodes that contain meta-information about sources; that is, information about the author, his or her position, intentions, and so forth (shaded boxes in Fig. 1). As can be seen in Fig. 1, the document nodes are only linked to central arguments in the situations model, which means that these arguments are mentally tagged for their source. Thereby, readers are able to take information such as the anticipated motive or the perceived expertise of an author into account when evaluating the reliability of an argument.

Britt et al. (1999) consider this model as "(...) typical of a good reader's model of multiple-text learning (...)" (p. 220), because information from different sources is represented in a highly integrated manner, while sources are separated from each other. However, empirical studies on the formation of documents models are rare, and their

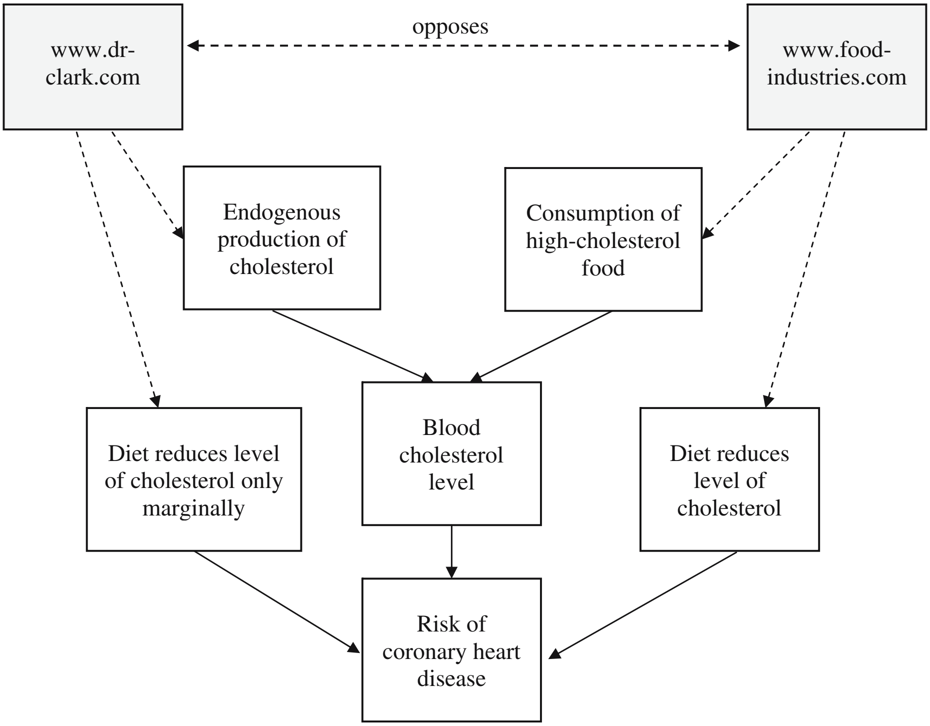

www.dr-clark.com
opposes
www.food-industries.com
Endogenous production of cholesterol
Consumption of high-cholesterol food
Diet reduces level of cholesterol only marginally
Blood cholesterol level
Diet reduces level of cholesterol
Risk of coronary heart disease

Fig. 1 Documents model of two documents written by different authors. The situations model is depicted as *boxes connected with solid lines*; the intertext model is depicted as *shaded boxes* that are linked (*dotted lines*) to selected arguments in the situation model

methods as well as findings are inconsistent. Britt et al. (1999) showed that readers can in fact form documents models when dealing with multiple texts. Undergraduate students were able to name the source of a given piece of information subsequent to reading a history text at a better than chance level. Yet, they did not mentally tag all information for their source, which is consistent with the documents model. Similarly, Rouet et al. (1996) found that college students showed some ability to integrate and relate information to sources. In their study, undergraduate history students integrated information from multiple documents revolving around a historical topic and organized it into a coherent essay text. Furthermore, these students were aware of the different status of different types of documents (e.g., historical essays vs. textbooks) and based their rankings of a document's trustworthiness on appropriate features such as the author's credentials or intentions. However, the results of Wineburg (1991) provide a more pessimistic view. Wineburg found that only expert history readers compared information across different sources and paid substantial attention to source information while dealing with the documents. High-school students did not attend to author information during reading and did not use author information to justify their credibility judgments provided after reading. In line with this rather pessimistic view, Britt and Aglinskas (2002) reported that the spontaneous use of source information when dealing with multiple documents in history was rather low both in college and high-school students.

Such inconsistencies reveal that one central question has yet to be answered sufficiently: Which factors determine whether readers actually form documents models? What leads them to integrate information and mentally tag contents for their sources when dealing with multiple documents? Up to now, empirical studies addressing these questions have focused on the role of task characteristics (Britt and Aglinskas 2002; Britt et al. 1999), features of the documents themselves (Britt et al. 1999), and the role of reader expertise (Rouet et al. 1997; Wineburg 1991). One of the main results supported by studies focusing on the role of task characteristics is that simple instructions to attend to source information are not sufficient to make readers deal with sources efficiently. Compared with readers receiving content instructions, readers receiving sourcing instructions neither performed better on a source identification task after reading (Britt and Aglinskas 2002), nor did they incorporate a larger amount of reliable information in a subsequent written essay (Britt et al. 1999).

Furthermore, expert-novice comparisons suggest an effect of expertise on dealing with sources in multiple documents situations. Wineburg (1991) reported that when confronted with a set of different history documents, history specialists qualified their choice of documents more accurately than novices did. Furthermore, specialists made extensive use of a metacognitive evaluation strategy called "sourcing heuristic," which involves attending to author information prior to reading a document. Novices, in contrast, applied this strategy only in a small number of cases.

However, Rouet et al. (1997) pointed out that in Wineburg's (1991) study, history specialists did not just differ from novices with regard to content expertise, but also with regard to the degree of discipline expertise at their disposal. In other words, through extensive training in dealing with different kinds of history documents, history specialists possess more sophisticated models of discourse structures within their discipline (Dillon 1991). This enables them to deal with multiple history documents more appropriately. In a comparison of graduate historians and graduate psychologists, Rouet et al. (1997) controlled for content expertise by choosing a historical topic unfamiliar to both groups. Results still showed significant differences between discipline experts and discipline novices. For instance, discipline experts were able to deal with the bias potentially included in participants' accounts. Furthermore, discipline experts tended to use multiple criteria when evaluating sources. Discipline novices, in contrast, based their evaluations mainly on content information and included less source information in their essays. The findings of Rouet et al. (1997) and Wineburg (1991) suggest that to fully understand which factors promote a successful processing of multiple documents, researchers need to address the concrete (meta-) cognitive strategies used by both expert and novice readers. This, however, has not been the focus of studies dealing with learning from multiple documents so far. With the present study, we seek to fill this void and shed some light on the role of metacognition in dealing with multiple documents on the WWW.

The role of metacognition in dealing with multiple documents on the WWW

The term metacognition is commonly referred to as the knowledge and regulation of cognition. It involves processes like planning, monitoring, evaluating, and elaborating (Baker and Brown 1984; Schraw and Moshman 1995). With regard to learning from texts, there is a large body of empirical evidence underlining the importance of metacognitive strategy use. When reviewing the literature pertinent to this topic, Baker and Brown (1984) concluded that proficient young readers monitor their ongoing comprehension and adapt their reading speed accordingly. Furthermore, they regularly activate prior knowledge and integrate new information into existing knowledge schemes. With the rise of hypermedia-based learning environments in educational contexts, the use of metacognitive strategies has become even more important. Due to their non-linearity, hypermedia-based learning environments afford a high amount of learner control, because laypersons have to make decisions on which information to access as well as the sequence in which to retrieve it (Dillon 2002; Dillon and Gabbard 1998). Furthermore, laypersons have to evaluate information in terms of its relevance to their current learning goal (Bannert 2003). Evidence for the importance of metacognition in dealing with multiple documents in hypermedia-based learning environments comes from intervention studies that systematically promote the use of metacognition (Bannert 2003; Lin and Lehman 1999). For instance, Bannert (2003) found that learning outcomes, as measured by a transfer test, were higher for students who received metacognitive prompts than for a control group.

We assume that metacognitive strategies are even more important when dealing with multiple documents on the WWW. The fact that the amount of immediately available information is nearly unlimited on the WWW underlines the need for a reasonable selection of information and a thorough self-monitoring of the comprehension process. Furthermore, laypersons need to activate prior knowledge in order to integrate information from multiple texts and thereby build semantic connections between information from different sources. Finally, to gain knowledge about the sources, laypersons have to evaluate sources in terms of quality and credibility. This involves finding out about the author as well as his or her credentials, intentions, possible affiliations, and sponsors.

However, in a study using think-aloud methodology, Stadtler and Bromme (2004) found that university students with little medical knowledge showed only moderate levels of metacognitive activity. Qualitative analyses of metacognitive activity further revealed that laypersons used inadequate criteria to judge the reliability of information provisions. They relied heavily on predictive judgments uttered before opening a web site as well as general impressions about the professionalism of a web site's layout uttered shortly after accessing a web site. Furthermore, laypersons rarely searched for author information or tried to find out about possible affiliations with commercial sponsors. This finding is in line with the results of Eysenbach and Köhler (2002). The authors report that adult laypersons were able to name adequate criteria for assessing a web site's reliability when explicitly asked to do so, but did not actually use them when conducting an Internet research on a medical topic.

Interestingly, in the study of Stadtler and Bromme (2004), use of the metacognitive strategies of monitoring, evaluating, and elaborating correlated significantly with knowledge acquisition. This result could be obtained for both the acquisition of factual knowledge as well as the comprehension of the subject matter. Moreover, the use of evaluation strategies related positively to the quality of essays on the credibility of sources. These results, although correlative in nature, point to the importance of metacognitive strategy use when dealing with multiple documents on the WWW.

This led us to develop the metacognitive tool *met.a.ware* (for a description of the tool, see the methods section), with which we sought to investigate the role of metacognition in dealing with multiple documents on the WWW. *Met.a.ware* encourages laypersons to monitor their comprehension and critically evaluate information by the means of metacognitive prompting. Metacognitive prompts focus the learners' attention toward their own cognition during the learning process (Brown 1997). The repeated prompting elicits metacognitive processes, which learners wouldn't show spontaneously. Evidence for the assumption that metacognitive prompts indeed impact on the metacognitive processes of learners has been found in studies using think-aloud methodology (Bannert 2004; Veenman et al. 1994). Thus, metacognitive prompting can be considered as particularly suitable in cases where learners are generally capable of executing metacognitive processes, but do not or only seldom apply these strategies spontaneously.

Predictions

We predicted that providing laypersons with monitoring prompts in *met.a.ware* would foster the acquisition of content knowledge (content knowledge hypothesis). We further predicted that providing laypersons with evaluation prompts in *met.a.ware* would foster the acquisition of knowledge about sources (source knowledge hypothesis), and that evaluation prompts would improve their ability to indicate the source of information after their Internet research (sourcing hypothesis). Finally, we predicted that laypersons receiving evaluation prompts would produce more arguments to justify their credibility judgments (justification of credibility rating hypothesis).

Method

Participants

A total of 80 undergraduate students at a German university (58 female) participated in the study.[1] Participants' age ranged from 19 to 32 with an average of 23.65 (SD=3.37). To ensure that participants were laypersons in the field of medicine, prior knowledge about the topic cholesterol was tested before the Internet search. One student scored more than 50% and was thus dropped from all further analyses. The remaining 79 participants scored an average of 4.61 (SD=2.47) out of 24 possible test points.

[1] Note that parts of the empirical research reported in this paper have been published in Stadtler and Bromme (2007), where further data on the effect of ontological classification are reported and *met.a.ware* is compared to a control group that took notes using paper and pencil.

Task and materials

Participants were confronted with a request by a fictitious friend. This friend had been diagnosed with a high level of cholesterol and now wants to make an informed decision on the question of whether to consent to medical treatment. Participants were asked to conduct Internet research in order to inform their friend about the topic cholesterol. For their Internet research, participants were provided with a set of 15 web sites that we had preselected. When selecting web sites, we took care that the resulting pool of information reflected the given heterogeneity of information available online in terms of information providers and their perspectives on this controversially discussed topic. Thus, we included web sites hosted by universities, nutritionists or journals in the field of medicine as well as companies from the food and pharmaceutical industries. Web sites were accessible via a list of links ordered alphabetically. They were displayed on a standard 17-in. computer screen and could be browsed using Microsoft Internet Explorer 6.

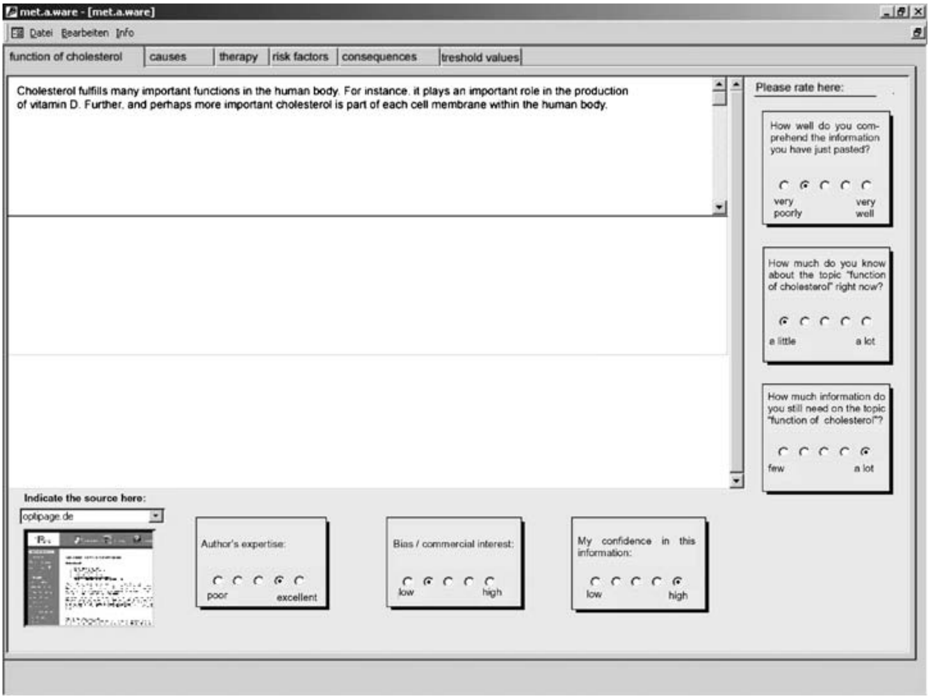
Development of the metacognitive tool *met.a.ware*

The computer-tool *met.a.ware* stimulates the use of metacognitive processes evaluation and monitoring. This is accomplished through the method of metacognitive prompting. As a monitoring prompt, laypersons are requested to assess how well they have comprehended the information they have just pasted, how much they currently know about the specific aspect of cholesterol, and how much information they still need regarding this aspect of cholesterol. They provide their answer by using 5-point rating scales (see right part of Fig. 2). As an evaluation prompt, laypersons are required to indicate the source of information each time they paste it into the *met.a.ware*. They also have to rate the author's credentials, the bias of information, as well as their confidence in the information on 5-point scales (see lower part of Fig. 2). Thus, evaluation prompts mainly focus laypersons on the source of a document. Ratings are attached permanently to the specific contents and can be retrieved by the user of *met.a.ware* at all times during future Internet research. Thus, laypersons add an additional layer of meta-information to the contents stored in *met.a.ware*.

Note that *met.a.ware* also provides laypersons with a means to store the information they have found on the WWW systematically. They do this by assigning information to different tabs labelled with aspects of, in this case, the topic cholesterol (ontological classification; see upper part of Fig. 2). The technical realization of *met.a.ware*, however, allows for a flexible adaptation of the tool towards other content domains, where different ontological categories and different types of prompts may be needed.

Design

Participants were randomly assigned to one of four groups that worked with different versions of *met.a.ware* or with a simple text window. To investigate the effects of metacognitive prompting we systematically varied the availability of prompts between the groups working with *met.a.ware*. Participants received either evaluation prompts (*evaluation* group), monitoring prompts (*monitoring* group), both types of prompts (*evaluation+monitoring* group). These conditions were compared with a group that did not receive metacognitive prompts (*no prompts* control group). All of the aforementioned conditions were provided with tabs for ontological classification and could copy and paste contents from the Internet into *met.a.ware*.



Cholesterol fulfills many important functions in the human body. For instance, it plays an important role in the production of vitamin D. Further, and perhaps more important cholesterol is part of each cell membrane within the human body.

Please rate here:

How well do you comprehend the information you have just pasted?
○ ● ○ ○ ○
very poorly … very well

How much do you know about the topic "function of cholesterol" right now?
● ○ ○ ○ ○
a little … a lot

How much information do you still need on the topic "function of cholesterol"?
○ ○ ○ ○ ●
few … a lot

Indicate the source here: optipage.de

Author's expertise:
○ ○ ○ ● ○
poor … excellent

Bias / commercial interest:
○ ● ○ ○ ○
low … high

My confidence in this information:
○ ○ ○ ○ ●
low … high

Fig. 2 Screenshot of the metacognitive tool *met.a.ware*

For the sake of completeness, we point out that to control for the effect of ontological classification, a second control group was introduced that is not described in this article. This group worked with a plain text window that allowed them to copy and paste information from the WWW, but provided neither ontological classification nor metacognitive prompts (*text window* control group). Results showed that the *text window* control group did not differ significantly from the *no prompts* control group on any of the dependent measures. Because the effect of ontological classification falls outside the scope of this article, results from the *text window* control group are not discussed any further (for a detailed picture of the results of ontological classification on Internet research, see Stadtler and Bromme 2007).

Likewise, we introduced a further control group that was only allowed to take notes using paper and pencil (*paper-and-pencil* control group). We thereby sought to investigate whether working with *met.a.ware* was superior compared to conducting one's Internet research without any external support through technical devices. Since the results on the comparison of *met.a.ware* with the *paper-and-pencil* control group fall outside of the scope of this article, they are not reported here (see Stadtler and Bromme 2007, for results of the *paper-and-pencil* control group).

Measures

Covariates

We collected data on *demographic variables* (four items), *computer and Internet experience* (four items), and *interest in the topic cholesterol* (four items) with a self-developed questionnaire. Moreover, we assessed participants' *need for cognition*, which is defined as the tendency to engage in and enjoy effortful cognitive endeavors (Cacioppo et al. 1984), with a German version of the original questionnaire devised by Bless et al. (1994). The measure comprises 16 Likert-type items and demonstrated good internal consistency (Cronbach's α=0.79). In addition, participants were required to indicate their subjectively perceived *time pressure* during Internet search using one Likert-type item.

Dependent variables

The formation of a documents model involves acquiring knowledge about contents and sources. Additionally, it requires a reader to mentally tag content information from the situations model to the respective source (Perfetti et al. 1999). Therefore, we developed two tests of content knowledge: a test of source knowledge and a measure of sourcing, as is described in the following sections.

Instruments measuring content knowledge

We had participants complete a self-developed 24-item multiple-choice test to measure their *factual knowledge* about the topic cholesterol. The measure's internal consistency proved to be good, as indicated by Cronbach's α=0.78. *Comprehension* of the subject matter was measured with four open questions, each requiring participants to compose a short written statement. The questions addressed central concepts of the subject matter, namely the risk-factor concept, the development of threshold values and the concept of relative and absolute risk reduction.

Instruments measuring source knowledge

Source knowledge was assessed with four items that were presented in a multiple-choice format and required participants to recall facts about the source of a web site. These included information crucial to a critical evaluation of a web site, i.e., the author's position, his or her affiliations, or the presence of commercial sponsors. The questions had to be answered for each web site visited during Internet search.

Sourcing

To examine to what degree laypersons tag information for their sources, participants were asked to write an argument-based essay on whether they thought it was worth trying to reduce cholesterol levels, and name the source of each argument they used.

Justification of credibility judgments

To measure participants' ability to justify their credibility judgments after Internet research, participants were requested to rate their three most preferred web sites in terms of credibility and subsequently give reasons for their judgments.

All measures were presented on-screen. Sample items for the measures used are given in the Appendix. Please note that we collected data on further variables (epistemological beliefs of participants), which we do not report in this paper, since they fall outside the scope of this article.

Procedure

Data collection was organized in a group setting with a maximum of seven participants at a time. We took care to ensure that participants worked individually, i.e., without interacting with other participants on the search task or on the completion of other assignments. Before conducting their Internet research, participants completed the 16-item questionnaire on demographic variables and covariates, i.e., computer and Internet experience, topical interest and need for cognition. In addition, participants' factual knowledge on the topic cholesterol was measured before the Internet research. Participants were then instructed on how to work with *met.a.ware*. We used a standardized video-instruction to inform participants about the features of *met.a.ware*, e.g., the meaning of the ontological categories as well as the function and value of the metacognitive strategies participants were intended to execute. Thereby, we sought to ensure that participants act in line with the metacognitive support provided (Bannert 2003).

After 40 min had elapsed, the experimenter requested participants to finish their Internet research. Search time was fixed in order to avoid time-on-task effects. Participants were additionally asked to rate the perceived time pressure after they had finished. After their Internet research, participants once again completed the multiple-choice test measuring factual knowledge and were requested to answer the four open questions measuring comprehension of the subject matter. They then wrote a short argument-based essay on whether they thought it was worth trying to reduce cholesterol levels, naming the source of each argument they used.

Additionally, knowledge about sources was assessed and participants were asked to rate the credibility of the three most appreciated web sites and to produce arguments to justify their ratings. Neither notes taken during the Internet research nor ratings provided in *met.a.ware* were available in the posttests. The whole session lasted about 100 min, on average.

Data analyses

Factual knowledge

We chose to calculate gain scores, i.e., the difference between factual knowledge posttest and pretest scores, because they provide a better interpretation of change between pretest and posttest than an analysis of covariance (ANCOVA) with prior knowledge as a covariable does (Rogosa 1988). Participants could score a maximum of 24 test points.

Comprehension scores

In a rating procedure, we scored the written answers to the four open comprehension questions in terms of soundness and detailedness. Participants could reach a maximum of 12 points on the four comprehension questions. To determine the procedure's reliability, two judges rated 10% of the answers blind to condition and independently from each other. Interrater-reliability as determined according to the formula of Holsti (1969) proved to be high, CR=94%.

Sourcing

For each argument in participants' essays on the question of whether it is worth trying to reduce high cholesterol levels, we determined whether participants named the correct source.

To obtain an index of sourcing, the number of correctly sourced arguments in participants' essays was related to the total number of arguments given.

Credibility judgments

Drawing on Wittwer et al. (2004), we developed a categorization scheme to analyze the number and type of arguments laypersons produced to justify their credibility judgments. Laypersons' arguments were classified using a set of mutually exclusive categories, which were called *Layout* (e.g., the professionalism, availability of pop-up ads), *Content* (e.g., internal consistency, agreement with information from other web sites), and *Source* (e.g., the author's expertise, her perceived motives). Inter-rater agreement for the coding process proved to be high, CR=95%.

Statistical analyses

We conducted planned contrasts between each of the experimental groups working with *met.a.ware* and the *no prompts* control group to test all a priori specified hypotheses in this paper. Thereby, we wanted to take a theory-driven approach, which entails the advantage of having a greater statistical power than post-hoc comparisons conducted in reaction to a significant omnibus *F*-test in an analysis of variance (Hays 1988; Rosenthal and Rosnow 2000). This is accomplished by reducing the probability that an existing effect is obscured by variation that is not of theoretical interest (Weinfield et al. 2000). Since planned contrasts do not require a significant omnibus *F*-test as a precondition, no omnibus *F*-tests are reported when planned contrasts were conducted (Czienskowski 1996). An alpha-level of 0.05 was chosen for all statistical tests unless otherwise indicated.

Results

Covariates

Separate ANOVAs were conducted for each of the four covariates Internet-/computer experience, interest in the topic, need for cognition and time pressure to find out whether there were any differences between groups on these variables. Because we did not expect to find any differences, an alpha-level of 0.20 was considered as statistically significant. However, none of the ANOVAs yielded a significant result (all Fs(3, 75)<1.64, ns) showing that groups did not differ on any of the covariates. As a consequence, the covariates were dropped from all further analyses.

Content knowledge

Factual knowledge

Mean pretest, post-test, and gain scores, as well as standard deviations for the four groups are presented in Table 1. Planned contrasts between each experimental group and the *no prompts* control group showed a significant difference between the *monitoring* group and the control group, $F(1, 75)=3.98$, $p=0.05$, $\eta^2_{part} = 0.05$.

Table 1 Mean pretest, post-test and gain scores for factual knowledge

Group	Pretest	Post-test	Gain scores
Monitoring	4.32 (1.91)	15.32 (2.36)	11.00 (3.04)
Evaluation+monitoring	4.80 (2.04)	14.75 (2.65)	9.95 (2.72)
Evaluation	4.30 (3.06)	14.50 (3.49)	10.20 (3.71)
No prompts control group	5.00 (2.75)	13.75 (3.71)	8.75 (4.36)

Standard deviations are given in brackets

As expected, no significant difference could be found between the *evaluation* group and the *no prompts* control group, $F(1, 75)=1.70$, $p=0.20$, $\eta^2_{part} = 0.02$. However, the planned contrast between the *evaluation+monitoring* group and the *no prompts* control group did not yield a significant difference either, which was not predicted by our hypothesis, $F(1, 75)=1.16$, $p=0.29$, $\eta^2_{part} = 0.02$.

Comprehension of the subject matter

Means and standard deviations with respect to the comprehension of the subject matter are shown in Table 2.

Consistent with our hypotheses, we did not find an effect of evaluation prompts on comprehension of the subject matter as shown by a non-significant contrast between the *evaluation* group and the *no prompts* control group, $F(1, 75)=0.28$, $p=0.60$, $\eta^2_{part} = 0.004$. However, contrary to our expectations, planned contrasts also failed to reveal any significant differences between the *monitoring* group and the *no prompts* control group, $F(1, 75)=1.71$, $p=0.20$, $\eta^2_{part} = 0.02$, and the *evaluation+monitoring* group and controls, $F(1, 75)=0.30$, $p=0.59$, $\eta^2_{part} = 0.004$.

Source knowledge

Table 3 depicts the mean percentage of correct answers on the source test. Percentages of correct answers were used instead of the total number of correct items, because participants were free to choose which web sites they visited. Thus, not all participants accessed all 15 web sites. To test our hypothesis that evaluation prompts would promote the acquisition of source knowledge, we performed planned contrasts between each of the three experimental groups receiving metacognitive prompts in *met.a.ware* and the *no prompts* control group. Both the *evaluation* group, $F(1, 75)=3.35$, $p=0.07$, $\eta^2_{part} = 0.04$, and the *evaluation+monitoring* group, $F(1, 75)=2.99$, $p=0.09$, $\eta^2_{part} = 0.04$, showed a trend towards better knowledge about source characteristics compared to the *no prompts* control group. As expected, no significant differences were found between the *monitoring* group and the *no prompts* control group, $F(1, 75)=1.03$, $p=0.31$, $\eta^2_{part} = 0.01$.

Table 2 Mean scores and standard deviations for comprehension

Group	M	SD
Evaluation	5.79	2.30
Monitoring	6.33	1.99
Evaluation+monitoring	5.80	2.15
No prompts control group	5.43	2.19

M Mean scores, *SD* standard deviation

Table 3 Mean percentage of items correct on the source test

Group	M	SD
Evaluation	45.33	10.80
Monitoring	33.83	12.60
Evaluation+monitoring	44.92	13.82
No prompts control group	37.97	13.43

M Mean percentage, *SD* standard deviation

Sourcing

Participants across the four conditions working with *met.a.ware* produced an average of 3.34 (SD=1.50) arguments in their essays on the question of whether to consent to medical treatment of the high cholesterol level. An ANOVA does not reveal any difference between conditions, $F(3,\ 75)=1.29$, $p=0.28$, $\eta^2_{part} = 0.05$. Means and standard deviations for the index of sourcing (i.e., the mean percentage of arguments that were tagged correctly for their source) are shown in Table 4. Given that, across all conditions, laypersons visited an average of 9.18 out the 15 pre-selected web sites, the average proportion of correctly sourced arguments (52.06%) was above chance in this sample. While the majority of laypersons did not tag each argument for its source (66%), there is also a considerable proportion of laypersons (34%) who were able to indicate the correct source for each argument they gave in their essay. Since neither the number of web sites visited, nor the number of arguments produced in the essays correlated significantly with the index of sourcing, these variables were not considered as covariates in subsequent planned comparisons.

Planned contrasts comparing each of the experimental groups receiving prompts with controls revealed a trend towards better source memory for the *evaluation* group, $F(1,\ 75)=3.51$, $p=0.07$, $\eta^2_{part} = 0.05$. Furthermore, laypersons in the *evaluation+monitoring* group significantly outperformed controls with respect to sourcing of arguments in their essays, $F(1,\ 75)=4.49$, $p=0.04$, $\eta^2_{part} = 0.06$. As expected, there was no significant difference when the *monitoring* group was compared with controls, $F(1,\ 75)=0.16$, $p=0.69$, $\eta^2_{part} = 0.002$. Therefore, the results support the hypothesis that evaluation prompts supported laypersons in mentally tagging content information for their sources.

Justification of credibility judgments

Using multivariate planned contrasts, each of the four experimental conditions working with *met.a.ware* was compared with the *no prompts* control group with respect to the number of arguments in each of the three categories *Content*, *Layout* and *Source*. As expected, the *monitoring* group did not differ significantly from the *no prompts* control group, $F(3,\ 73)=$

Table 4 Mean percentage of correctly sourced arguments

Group	M	SD
Evaluation	62.17	39.15
Monitoring	42.72	44.28
Evaluation+monitoring	65.42	42.86
No prompts control group	37.46	40.55

M Mean percentage, *SD* standard deviation

0.62, p=0.60, $\eta^2_{part} = 0.002$. However, we found a marginally significant difference between the *evaluation* group and controls on the multivariate level, F(3, 73)=2.21, p= 0.09, $\eta^2_{part} = 0.08$. Contrary to our expectation, the *evaluation+monitoring* group did not differ significantly from controls, F(3, 73)=1.76, p=0.16, $\eta^2_{part} = 0.07$.

Univariate contrasts revealed that the multivariate effect in the evaluation condition can be attributed to a significant difference between the evaluation group and the *no prompts* control group with respect to the number of arguments in the category *Source*, F(1, 75)= 4.71, p=0.03, $\eta^2_{part} = 0.06$. Likewise, members of the *evaluation+monitoring* showed a tendency towards more arguments classified as belonging to the category *Source* compared with the *no prompts* control group, F(1, 75)=3.27, p=0.07, $\eta^2_{part} = 0.04$.

As expected, no significant differences could be found between the monitoring group and the *no prompts* control group with respect to the number of arguments in the three categories *Source, Content and Layout*, all Fs(1, 75)<1.53, ns. Also, none of the planned comparisons between the experimental groups receiving metacognitive prompts and the controls with respect to the categories *Content* and *Layout* yielded any significant differences, all Fs(1, 75)<1.40, *ns*.

Thus, the results confirm our hypothesis partially. Laypersons in both conditions that received evaluation prompts produced more arguments focusing on the author of a web site than controls. However, they did not produce more arguments with regard to content and the web site's layout.

Discussion

With our present study, we wanted to investigate the role of metacognition in dealing with multiple documents on the WWW. More precisely, we sought to examine whether prompting for the metacognitive processes monitoring and evaluation would help laypersons to form documents models, i.e., to gain knowledge about contents, sources and to mentally tag content information for their source (cfr, Perfetti et al 1999).

The results with respect to the acquisition of factual knowledge partially support our hypothesis. Compared with the *no prompts* control group, participants receiving monitoring prompts acquired significantly more factual knowledge on the topic cholesterol. Here, the repeated prompts to monitor one's comprehension of the material pasted into *met.a.ware* as well as the prompt to determine one's information needs successfully fostered laypersons' formation of a content representation. This is line with our previous research (Stadtler and Bromme 2004) where we could show that spontaneous comprehension monitoring during Internet research was significantly correlated with the acquisition of factual knowledge. Research on text comprehension has suggested that a thorough self-monitoring is crucial for learning from text (Baker and Brown 1984). We argued that these skills become even more important during Internet research, since laypersons are confronted with unmanageable masses of information that are spread across multiple documents, which are sometimes only loosely connected. With the results from the *monitoring* group, we were able to back up this claim with empirical data showing an advantage of an increased self-monitoring during Internet research.

Still it needs clarification why laypersons from the *evaluation+monitoring* group did not differ significantly from controls in their acquisition of factual knowledge. One possible explanation is that the requirement to react to both types of prompts each time they pasted information into *met.a.ware* may have been too demanding for laypersons. It is conceivable that due to the requirements of *met.a.ware*, laypersons in the *evaluation+monitoring* condition may have had less time to elaborate and memorize information than laypersons in the *monitoring* group who only had to react to one type of prompt. However, this explanation is not supported by data we collected on the subjectively perceived time pressure during Internet research. Here, we could not find a difference between the *monitoring* group and the *monitoring+evaluation* group. Another explanation why we did not find greater differences in knowledge acquisition between groups receiving monitoring prompts and the *no prompts* control group might lie in the choice of the control group itself. One should bear in mind that in the *no prompts* control group, laypersons were provided with a copy-and-paste tool that provided structure through ontological classification and different text slots. This might have had a supportive effect on laypersons' Internet research that obscured the effects of metacognitive prompting. The results of Stadtler and Bromme (2007) support this notion. Here the groups receiving metacognitive prompts were compared with another control group that was only allowed to take notes using paper and pencil. Results revealed that both the *monitoring* group and the *monitoring+evaluation* group significantly outperformed laypersons who conducted their Internet research without the assistance of external support.

Apart from measuring factual knowledge, we collected data on the comprehension of the subject matter. The results failed to reveal an improved performance of the conditions receiving comprehension prompts compared to the *no prompts* control group. In what follows, we offer two competing explanations for this state of affairs. First, our findings may be due to the fact that developing a deep-level understanding of contents within the time frame of 40 min was a highly challenging task for participants who had low prior topical knowledge. Results of previous research (Stadtler 2006) have shown that when laypersons were confronted with a similar scenario, they first tried to gather some factual knowledge, such as what are threshold values for cholesterol or which diseases may result as a consequence of too high levels of cholesterol. After they had learned about these facts, they were willing to tackle more complex issues such as the interplay between cholesterol and other risk factors for developing coronary heart disease. These practical constraints may well explain why the mean scores on the comprehension task in the present study were rather low in all groups and we were not able to detect group differences.

Alternatively, the results may rather reflect the specific 'learning goal' laypersons pursue during Internet research. As Bromme et al. (2005) pointed out, laypersons are not novices, i.e., they do not want to become experts in the area of their inquiry. As a consequence, they may be satisfied with a metonymic, i.e., partial understanding of concepts. This may include basic knowledge about facts as measured in the factual knowledge test, but not a deeper understanding of more complex issues such as the interplay of different risk factors for developing coronary heart disease, the origins of threshold values or the difference between relative and absolute risk reduction. If this explanation accounts for our current findings, it would be unlikely that enhancing the search time would result in deeper understanding of the subject matter. We would rather expect that laypersons finish their web search process after a subjectively sufficient level of understanding had been achieved. Further research is needed to address this issue and clarify which explanation accounts for the current findings.

However, forming a full documents model does not just require knowledge about contents, but also a representation of knowledge about sources (Perfetti et al. 1999). This is particularly crucial when dealing with medical information on the WWW, because single documents may contain faulty or biased information and not always provide a reliable account. This is why we gave laypersons evaluation prompts requiring them to rate information in terms of its credibility. Results on testing source knowledge revealed that the intensified dealing with the sources of information improved performance: Members of both the *evaluation* group and the *evaluation+monitoring* group showed a tendency to recall more information about sources than controls. This underlines the importance of metacognitive strategies in the formation of source knowledge as well. Because most laypersons do not routinely employ evaluation strategies, such as identifying the author of a document prior to reading it, instructional support is needed to let laypersons gain knowledge of contents in addition to their representation of contents.

The results further show that prompting laypersons to evaluate information enhances their ability to produce arguments to justify their credibility ratings. However, this effect was only observed for arguments relating to the source of a document. No differences between conditions were obtained with respect to the number of arguments relating to the quality of information or the layout of the web site. Taking into account that the evaluation prompts mainly focused laypersons' attention to the source of a document, this is a plausible result. The prompting procedure, however, did not trigger comprehensive processes of information evaluation but impacted in a more specific way on evaluation activities. One explanation for this finding is that our attempt to focus laypersons on the evaluation of information competed with a 'content focus' induced by the search task itself. Laypersons were instructed to conduct Internet research to inform a fictitious friend that had been diagnosed with a high level of cholesterol, which is a challenging task for them. Evaluating information might have been perceived as an additional challenging demand so that laypersons restricted themselves to evaluate the source of information and did not to engage in further evaluation activities involving other criteria, such as the quality of information or the web site's layout.

Further research is needed to determine whether prompts that focus laypersons on other aspects of credibility, such as the internal consistency of information or its consistency with information found on further web sites, would enhance laypersons' ability to produce arguments to justify credibility judgments in a similar way. In addition, it should be examined whether it is possible to focus laypersons on more than one dimension of evaluation without impairing the formation of a sound representation of contents.

Finally, we found an effect of evaluating information on the sourcing of information. Laypersons who received evaluation prompts were better able to indicate the source of their arguments in an essay task after Internet research. Obviously, the intensified dealing with the sources of information during Internet research made sources salient and led laypersons to create a stronger mental link between contents and sources. This enabled laypersons to weigh up their arguments with respect to the question of whether to reduce one's cholesterol level in the light of the authors' motives, his or her perceived expertise, or the perceived bias of information.

The fact that the majority of laypersons did not tag each argument for its source is consistent with previous research on sourcing when dealing with multiple documents (Britt et al. 1999). Given the high cognitive demands of learning from multiple documents on the Internet, selectively tagging only the most important information for its source is a reasonable strategy and consistent with the assumptions of the documents model (Perfetti et al. 1999). Still, we found a comparably high number of laypersons who were able to correctly indicate the source of each argument they gave in their essays. This result may be due to our methodology in which laypersons were asked to indicate the source of arguments that they deemed important enough to be included in their essays. In the terminology of Perfetti et al. (1999), these were "core arguments" chosen by the laypersons themselves. It is likely that the proportion of correctly sourced arguments would have been lower if the measure of sourcing had been ascertained through an experimenter-directed presentation of stimuli that contained both core and non-core arguments. However, the present way of assessing the degree of sourcing in laypersons bears the advantage that it measures the degree of sourcing in an applied context, i.e., where laypersons directly make use of the mental connection between contents and sources. Summing up, the results reveal that the integration of source information and content information while dealing with multiple sources on the Internet is not only a desideratum but a realistic goal that can be fostered through the metacognitive strategy of evaluating information.

Taken together, this study provides evidence that the use of metacognition plays an important role in the formation of documents models when dealing with multiple documents on the Internet. Stimulating evaluation processes through metacognitive prompting successfully fostered the formation of the intertext model: laypersons acquired knowledge about sources and showed better tagging of content information for their sources. Moreover, laypersons were able to apply their knowledge about sources when justifying their credibility judgments. Further studies are needed to examine the conditions under which the stimulation of monitoring processes improves the formation of the situations model, as results concerning the acquisition of content knowledge were less conclusive. The results also have practical implications as they open up the possibility of designing intervention programs to support laypersons in dealing with multiple documents on the WWW by fostering the use of metacognitive strategies.

Acknowledgements The research reported in this paper has been supported by a scholarship from the *Deutsche Forschungsgemeinschaft* (DFG). The scholarship has been granted to the first author in the context of the Virtual Ph.D. Program "Knowledge Acquisition and Knowledge Exchange with New Media."

Appendix

Sample question for the multiple-choice test on factual knowledge:

For what purpose does our body need cholesterol?

- To transport oxygen in the blood
- To build cell membranes
- To break down carbohydrates
- To synthesize vitamin C
- Our body doesn't need cholesterol
- I don't know

Note: Each of the 24 items consisted of four distractors, one attractor and one "I don't know" option, which was included to reduce the effect of guessing.

Sample questions for the open comprehension questions:

- Why do some researchers criticize the reduction of threshold values for the blood cholesterol level?
- Is it reasonable to assess the individual risk for coronary heart disease solely on the basis of the blood cholesterol level?

Sample questions for the need for cognition questionnaire:

- "I really enjoy a task that involves coming up with new solutions to problems."
- "I would prefer complex to simple problems"

Note: Agreement was rated on a 7-point-Likert scale, in which 1 was labeled "totally agree" and 7 was labeled "totally disagree".

Sample questions for the multiple-choice test on knowledge about source information What is the profession of the author of the information on this web site?

- Physician
- Scientist
- Nutrionist
- Journalist
- Layperson
- There is no information about the author available on the web site
- I don't know

Is there any advertisement for a cholesterol-related product on the web site?

- Yes
- No
- I don't know

References

Baker, L., & Brown, A. L. (1984). Metacognitive skills and reading. In P. D. Pearson (Ed.), *Handbook of reading research*. New York: Longman.

Bannert, M. (2003). Effekte metakognitiver Lernhilfen auf den Wissenserwerb in vernetzten Lernumgebungen [Effects of metacognitive learning aids in networked learning environments]. *Zeitschrift für Pädagogische Psychologie, 17*, 13–25.

Bannert, M. (2004). Designing metacognitive support for hypermedia learning. In H. Niegemann, D. Leutner & R. Brünken (Eds.), *Instructional Design for Multimedia-Learning*. (pp. 19–30). Münster: Waxmann.

Bhavnani, S. K., Jacob, R. T., Nardine, J., & Peck, F. A. (2003). *Exploring the distribution of online healthcare information.* Paper presented at the CHI, Fort Lauderdale, FL, USA (April).

Bless, H., Wänke, M., Bohner, G., Fellhauer, R. F., & Schwarz, N. (1994). Need for cognition: Eine Skala zur Erfassung von Engagement und Freude bei Denkaufgaben. [Need for cognition: A scale measuring engagement and happiness in cognitive tasks]. *Zeitschrift für Sozialpsychologie, 25*, 147–154.

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. v. d. Broek (Eds.), *Narrative comprehension, causality, and coherence: Essays in honor of Tom Trabasso*. Mahwah, NJ: Erlbaum.

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

Bromme, R., Jucks, R., & Runde, A. (2005). Barriers and biases in computer-mediated expert-layperson-communication. In R. Bromme, F. W. Hesse, & H. Spada (Eds.), *Barriers, biases and opportunities of communication and cooperation with computers—and how they may be overcome*. New York: Springer.

Brown, A. L. (1997). Transforming schools into communities of thinking and learning about serious matters. *American Psychologist, 52*, 399–413.

Cacioppo, J. T., Petty, R. E., & Chuan, F. K. (1984). The efficient assessment of need for cognition. *Journal of Personality Assessment, 48*, 306–307.

Czienskowski, U. (1996). *Wissenschaftliche Experimente: Planung, Auswertung, Interpretation*. [Scientific experiments: planning, analysis and interpretation]. Weinheim: Psychologie Verlags Union.

Dillon, A. (1991). Readers' models of text structures: The case of academic articles. *International Journal of Man–Machine Studies, 35*, 913–925.

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*. New York: Pergamon.

Dillon, A., & Gabbard, R. (1998). Hypermedia as an educational technology: A review of the quantitative research literature on learner comprehension, control, and style. *Review of Educational Research, 68*, 322–349.

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

Eysenbach, G., Powell, J., Kuss, O., & Eun-Ryoung, S. (2002). Emprical studies assessing the quality of health information for consumers on the World Wide Web. A systematic review. *JAMA, 287*, 2691–2700.

Goldman, S. R., & Rakestraw, J. A. (2000). Structural aspects of constructing meaning from text. In M. L. Kamil, P. Mosenthal, P. D. Pearson, & R. Barr (Eds.), *Handbook of reading research, vol. III* (pp. 311–335). Mahwah, NJ: Erlbaum.

Hays, W. L. (1988). *Statistics*, 4th edition. Fort Worth, TX: Holt, Rinehart & Winston Inc.

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

Holsti, O. R. (1969). *Content analysis for the social sciences and humanities*. USA: Addison-Wesley.

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

Kintsch, W., & van Dijk, T. A. (1978). Toward a model of text comprehension and production. *Psychological Review, 85*, 363–394.

Lin, X., & Lehman, J. D. (1999). Supporting learning of variable control in a computer-based biology environment: Effects of prompting college students to reflect on their own thinking. *Journal of Research in Science Teaching, 36*, 837–858.

O'Connor, A. M. (1995). Validation of a decisional conflict scale. *Medical decision making, 15*, 15–30.

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.

Rogosa, D. (1988). Myths about longitudinal research. In: K. W. Schaie, R. T. Campbell, W. M. Meredith, & S. C. Rawlings (Eds.), *Methodological issues in aging research* (pp. 171–209). New York: Springer.

Rosenthal, R., & Rosnow, R. L. (2000). *Contrasts and effect sizes in behavioral research: A correlational approach.* Cambridge, New York: Cambridge University Press.

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.

Schraw, G., & Moshman, D. (1995). Metacognitive theories. *Educational Psychology Review, 7*, 351–371.

Stadtler, M. (2006). *Auf der Suche nach medizinischen Fachinformationen. Metakognitionen bei der Internetrecherche von Laien* [Searching for medical information. The role of metacognition in the Internet research of laypersons]. Münster: Waxmann.

Stadtler, M., & Bromme, R. (2004). Laypersons searching for medical information on the web: The role of metacognition. In: K. Forbus, D. Gentner, & T. Regier (Eds.), *Proceedings of the 26th Annual Conference of the Cognitive Science Society* (p. 1638). Mahwah, NJ: Erlbaum.

Stadtler, M., & Bromme, R. (2007). Effects of the metacognitive computer-tool *met.a.ware* on the web search of laypersons. *Computers in Human Behaviour* (in press).

Veenmann, M. V., Elshout, J. J., & Busato, V. V. (1994). Metacognitive Mediation in learning with computerbased simulations. *Computers in Human Behavior, 10*, 93–106.

Weinfield, N. S., Sroufe, L. A., & Egeland, B. (2000). Attachment from infancy to early adulthood in a high-risk sample: Continuity, discontinuity, and their correlates. *Child Development, 71*, 695–702.

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.

Wittwer, J., Bromme, R., & Jucks, R. (2004). Kann man dem Internet trauen, wenn es um die Gesundheit geht? Die Glaubwürdigkeitsbeurteilung medizinischer Fachinformationen im Internet durch laien. [Is the Internet a trustworthy source when it comes to health information? Laypersons' trustworthiness ratings of medical information on the Internet] *Zeitschrift für Medienpsychologie, 2*, 48–56.