Source evaluation of domain experts and novices during Web search

Source Evaluation of Domain Experts and Novices during Web Search

Saskia Brand-Gruwel [a], Yvonne Kammerer [b], Ludo van Meeuwen [a], & Tamara van Gog [c,d]

[a] Welten Institute – Research Centre for Learning, Teaching, and Technology, Open University of the Netherlands, Heerlen, The Netherlands

[b] Knowledge Media Research Center, Tuebingen, Germany

[c] Department of Psychology, Education, and Child Studies, Erasmus University Rotterdam

[d] Department of Education, Utrecht University

Correspondence concerning this paper should be addressed to Saskia Brand-Gruwel, Open University of the Netherlands, PO Box 2960, 6401 DL Heerlen, The Netherlands. E-mail: saskia.brand-gruwel@ou.nl

Abstract

Nowadays, almost everyone uses the World Wide Web (WWW) to search for information of any kind. In education, students frequently use the WWW for selecting information to accomplish assignments such as writing an essay or preparing a presentation. The evaluation of sources and information is an important sub-skill in this process. But many students have not yet optimally developed this skill. On the basis of verbal reports, eye-tracking data, and navigation logs this study investigated how novices in the domain of psychology evaluate Internet sources as compared to domain experts. In addition, two different verbal reporting techniques, namely thinking aloud and cued retrospective reporting, were compared in order to examine students' evaluation behavior. Results revealed that domain expertise has an impact on individuals' evaluation behavior during Web search, such that domain experts showed a more sophisticated use of evaluation criteria to judge the reliability of sources and information and selected more reliable information than domain novices. Furthermore, the different verbal reporting techniques did not lead to different conclusions on criteria use in relation to domain expertise, although in general more utterances concerning evaluation of sources and information were expressed during cued retrospective reporting.

Keywords: web search, source evaluation, thinking aloud, cued retrospective report

1. Introduction

The World Wide Web (hereafter referred to as the Web) has evolved into a major information resource offering easy access to billions of Websites on almost any topic. Accordingly, almost everyone uses the Web to search for a variety of information nowadays, and students frequently use search engines to find information that will help them accomplish assignments such as writing an essay or preparing a presentation (e.g. Metzger, Flanagin, & Zwarun, 2003; Purcell, Heaps, Buchanan, & Friedrich, 2013). Similarly, scholars frequently use the WWW to search for information regarding their research area, with search engines being a popular discovery tool (Jamali & Asadi, 2010). However, the process of information search (also known as information seeking), is not always optimal. It can be regarded as a problem-solving process driven by an information problem (e.g., Brand-Gruwel, Walraven, & Wopereis, 2009; Marchionini, 1995; Kuhlthau, 1993; Wilson, 1999; Wopereis, Brand-Gruwel, & Vermetten, 2008), and as for other problem-solving skills, this means that students need to master a certain number of sub-skills in order to perform well on the information search tasks.

An information problem is a problem that requires new information, which must be searched for, in order to solve it. An information problem "arises when a discrepancy occurs between information needed to answer a certain question and information already known" (Walraven, Brand-Gruwel, & Boshuizen, 2009, p. 235). Being able to solve information problems means that individuals are able to identify information needs, locate information sources, judge these sources (e.g., in terms of quality, trustworthiness, etc.), extract and organize the right information from each source, and then synthesize information from a variety of sources (Brand-Gruwel, Wopereis & Vermetten, 2005). On the Web, the amount of information is enormous and there are hardly any gatekeepers that filter information. Therefore, when

searching the Web, the critical evaluation of sources and information is a very important sub-skill in information problem solving (Brand-Gruwel & Stadtler, 2011; Kammerer & Gerjets, 2012; Rouet, 2006). This is particularly true when the information problems concern complex scientific topics, for which there may not be any clear-cut answers. This results in a large amount of disagreement across sources, and students have to be able to deal with this uncertainty when evaluating information (sources) on such topics (Bråten, Strømsø, & Salmerón, 2011; Kobayashi, 2009). Various studies have addressed the evaluation behavior of students when dealing with multiple documents and conflicting information (e.g., Braasch, Bråten, Strømsø, Anmarkrud, & Ferguson, 2013; Braasch, Rouet, Vibert, & Britt, 2012; Lawless, Goldman, Gomez,Manning, & Braasch, 2012; Rouet, Britt, Mason, & Perfetti, 1996; Strømsø, Bråten, Britt, & Ferguson, 2013; Tabatabai, & More, 2005; Van Strien, Brand-Gruwel, & Boshuizen, 2014; Walhout, Brand-Gruwel, Van Dijk, Jarodzka, De Groot, & Kirschner, 2015). It can be concluded that dealing with conflicting information when solving information problems is a complex process, and students have difficulties with knowing which criteria to use when evaluating, knowing were to look on websites to find source information and how to interpret that source information. Also it can be stated that research on spontaneous evaluation behaviour is scarce. Furthermore, students' domain specific epistemic beliefs, attitude and prior knowledge are factors that are known to influence this process.

Because prior knowledge about a domain can facilitate the evaluation of information and information sources (as will be discussed in more detail below), the present study examines the role of domain expertise in source evaluation during Web search by analyzing verbal protocols, eye-tracking data, and navigation logs. Specifically, we aim to gain more detailed insight in the criteria domain novices and domain experts use to evaluate Web sources and the information

therein, when selecting Websites that can be used as a basis for writing an article about a psychology topic for a popular psychology magazine. This also allows for examining whether the use of more sophisticated evaluation criteria is related to the selection of more trustworthy Websites. It can be expected that using sophisticated criteria and thinking more critically about the trustworthiness and reliability of Websites will lead to a selection of more trustworthy Websites and information.

In addition, a second research aim is to compare two different verbal reporting techniques on the evaluation behavior of domain experts and novices, to identify if different methods come up with different results. A method often used in research on the use of evaluation criteria is concurrent thinking aloud (e.g. Brand-Gruwel, Walraven, & Wopereis, 2009). However, for instance the study of Walraven, Brand-Gruwel, and Boshuizen (2009) showed that participants skipped sites and information without expressing why while thinking aloud. This may be associated with high cognitive load novices experience during the search process, which makes it hard for them to continue to think aloud (cf. Van Gog, 2006). Experts on the other hand, might have automated evaluations to such an extent that they also fail to report them during thinking aloud. This raises the question of whether participants did not actually evaluate more information than reported, or whether they did make more evaluations but failed to verbally express them. In case of the latter, the cued retrospective reporting method, in which participants report their thoughts after accomplishing a task, while looking at a replay of their own task performance with their eye-movements overlaid (Van Gog, Paas, Van Merriënboer, & Witte, 2005), is a promising method that might elicit more utterances about evaluation criteria used during Web search. For novices this might be the case because the cognitive load of the task no longer limits their report (in the study by Van Gog et al., 2005, they preferred cued retrospective reporting to thinking

aloud, as reported in Van Gog, 2006), whereas for experts, the cues provided by the replay might make them aware of (and report on) the criteria they (automatically) applied while searching for information. In the present study the two verbal reporting techniques will be compared, in order to examine whether the information that is gained about evaluation behavior of domain experts and novices differs using the two techniques.

1.1 Source evaluation behaviour and domain expertise

More than two decades ago, research on the reading of historical texts has pointed out that experts in this domain regularly paid attention to source information (e.g., a document's author) during historical problem solving. In his seminal study, Wineburg (1991) investigated expert historians and high-school students (i.e., novices) reading multiple documents 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, the novice students often tended to ignore source information altogether. In the same vein, Rouet, Favart, Britt, and Perfetti (1997) found that novices in the field of history (i.e., graduate students in psychology) mainly focused on a document's content to evaluate its usefulness when reading multiple documents on controversial historical events, whereas students with higher expertise in the field of history (i.e., graduate students in history) also considered the type of source (e.g., a textbook, reports by contemporary witness, historian essays). Likewise, a thinking-aloud study by Wyatt et al. (1993) showed that social scientists who were asked to read self-selected scientific articles in their field of research constantly monitored the credibility of the texts they were reading. A study by Stadtler, Scharrer, Brummernhenrich, and Bromme (2013) on how medical experts and laypeople deal with conflicting information from science

texts reveal that experts report conflicts in the multiple documents more often than the laypeople. In general though (i.e. for both experts and novices), multiple document reading stimulated the integration of conflicting information.

With regard to Website evaluation, a survey study by Stanford, Tauber, Fogg, and Marable (2002) examined which criteria and features domain experts and novices reportedly used to evaluate the credibility of health and finance Websites. Whereas domain novices based their credibility evaluations mainly on the Website design (e.g., colours, 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. Navigation logs recorded in a Web search study by Hölscher and Strube (2000) revealed that domain experts (university students of economics) spent significantly less time on each Website that was selected from a search engine results page (SERP) in order to solve a series of information problems than domain novices (university students of other subjects). The authors argued that this likely resulted from the fact that domain experts need less time to read the information and can also decide faster how to move on.

Other studies did not compare domain novices and domain experts but investigated differences between groups of relatively novice students who varied in their level of prior domain knowledge. For instance, Bråten, Strømsø, and Salmerón (2011) found that when reading multiple documents about climate change, undergraduates with little knowledge of the subject matter trusted different documents to the same extent, irrespective of the type of source, whereas students with higher prior knowledge judged an article issued by a company with vested interests in the addressed issue as being less trustworthy than other documents. During Web search, when users themselves are responsible for selecting a manageable subset of useful information sources

for further study, students with moderate domain knowledge have been shown to scrutinize search results presented by a search engine more thoroughly, by examining the titles, the page excerpts, and the URLs of the search results, than low-knowledge students, as indicated by retrospective interviews (MaKinster, Beghetto, & Plucker, 2002) or eye-tracking data (Kammerer & Gerjets, 2013). When deciding which Websites to use for further study by bookmarking them, students with higher domain knowledge reported that they bookmarked Websites more often on the basis of the trustworthiness of the site, and less often on the basis of link position in the search engine results page (SERP), as compared to those with little domain knowledge (Salmerón, Kammerer, & García-Carrión, 2013). This pattern of results suggests that individuals with higher domain knowledge relied on deep rather than on superficial cues in their assessment of Websites.

Another body of research looked into the difference concerning source evaluation between better and poorer learners. The study by Goldman, Braasch, Wiley, Graesser and Brodowinska (2012) revealed that better undergraduates learners engaged more in sense-making, self-explanation and comprehension-monitoring processing on reliable websites as compare to unreliable websites, and did this more often than the poorer learners. Moreover, research is showing a relation between source evaluation behavior and learning outcomes. Wiley, Goldman, Graesser, Sanchez, Ash, and Hemmerich (2009) investigated how undergraduates studying Internet sources to find causes for the Mt. St. Helens eruption, and found evidence that source evaluation significantly predicts learning outcomes. Moreover, their findings indicated that successful learners were better able to discriminate between scientific reliable and unreliable information.

In the present study, we aim at expanding this line of research by examining differences in information processing and source evaluation between domain novices and domain experts (cf. Wineburg, 1991), when being instructed to select a subset of Websites presented by a search engine that can be used as basis to write an article about a conflicting psychology topic. For this purpose we use a rich multi-method approach including verbal protocols, eye-tracking data and navigation logs as process measures as well as the final selection of Websites as outcome measure.

1.2 Verbal reporting techniques

In order to design instruction to foster students' information problem solving skills, gaining more insight into the cognitive processes involved when individuals search the Web for information and the way they evaluate the usefulness and reliability of Websites and information is important. For this purpose, participants are often instructed to think aloud during their search, and such think-aloud (TA) protocols are used to gain information about the information participants are focusing on, and the criteria they use in evaluating it and selecting (or dismissing) it for later use (e.g. Brand-Gruwel et al., 2005; Gerjets, Kammerer, & Werner, 2011; Mason, Ariasi, & Boldrin, 2011). This concurrent reporting technique that requires participants to verbalize all thoughts that come to mind during task performance is a widely used verbal technique to uncover cognitive and metacognitive processes. However, a disadvantage of concurrent reporting is that it may become difficult to maintain with novice learners or with highly complex tasks due to high cognitive load (Van Gog et al., 2005; Van Gog, Kester, Nievelstein, Giesbers, & Paas, 2009).

An alternative for concurrent reporting that can overcome the problem of cognitive load is retrospective reporting. Retrospective reporting requires learners to report the thoughts they had while they were working on a task immediately after task performance (Conrad, Blair, & Tracy, 1999; Ericsson & Simon, 1993; Van Someren, Barnard, & Sandberg, 1994). However, retrospective reports have the drawback –especially on tasks that require more than a few minutes to complete- that information is omitted (Van Gog et al., 2005; Van Gog et al., 2009). For instance, Kuusela and Paul (2000) reported that concurrent protocols contained more information on actions than retrospective protocols, arguing that retrospective reports often only contained references to the effective actions that led to the completion of the task. Furthermore retrospective reporting, is sensitive to fabrication, that is, reporting of actions that were not actually taken.

A specific form of retrospective reporting that could counteract these concerns, is Cued Retrospective Reporting (CRR; Van Gog et al., 2005). CRR is a method to capture verbalizations of thought processes after task performance cued by a replay of a recording of the problem-solving process, with one's own eye movements overlaid. Eye-tracking equipment nowadays allows not only for recording, but also for replaying eye movements overlaid on a recording of the computer screen and all actions performed on that screen, which can cue participants' memory of their thoughts during the task performance process, reducing the risk of omissions and fabrications. Especially in web-search tasks, the fact that participants do not only see what they typed in or clicked on during their search, as they would in a normal screen recording, but also see what information they attended to, is potentially very useful (e.g., they might attend to several search results before selecting one, meaning they engaged in some kind of evaluation that

would not be apparent from a normal screen recording, which would only show a mouse click on the selected result).

In a direct comparison of these three verbal reporting techniques, Van Gog et al. (2005) found that CRR and TA resulted in *quantitatively* more information being reported than did retrospective reporting. However, they also note that even though CRR and TA both performed well on a quantitative level, the question of whether there are *qualitative* differences in the information uncovered with those techniques should be further investigated, as should the question of whether the techniques are differentially effective for different groups of participants (e.g., different levels of expertise). In the present study we address these questions by comparing TA and CRR in terms of the data they provide concerning how people search the Web and evaluate information and sources found on the Web.

1.3 The present study

In the present study, domain novices and experts in Psychology were presented with a search engine results page (SERP) from which they were instructed to select Websites that they felt could be used as a basis for writing an article for a popular psychology magazine. The first research question addressed in this study concerns the differences between domain novices and experts in their evaluation behaviour (evidence by the verbal reports, log files, and eye movement data) when evaluating search engine results and the quality of the sources they select. We address this question using data from two tasks performed under two different verbal reporting techniques (which task was performed under which condition was counterbalanced across participants), CRR and TA. This also allows for addressing our second research question

of what the differences are between these techniques in terms of the data they yield concerning the first research question.

Taken the above-described literature, it is hypothesized that domain experts would express less superficial utterances concerning the reliability and usefulness of sources and information (H1) and more specific utterances on reliability evaluation than the domain novices (H2). Furthermore, it is expected that domain experts would fixate (i.e., look at) on more search results before accessing the first Website from the SERP, which would indicate more careful consideration of which Websites to access (H3), are more likely to attend to source information in deciding whether a Website should be selected for use in writing the article (H4), and spend less time on a Website to decide that it should be selected then domain novices (H5). Concerning the performance outcome measure (i.e., selection of websites for writing the article) it is expected that domain experts would select more trustworthy Websites than domain novices (H6). Whether or not TA and CRR would lead to different insights regarding Hypotheses 1 and 2, is explored as an open question.

2. Method

2.1 Participants

The group of domain novices consisted of nineteen students who volunteered to participate in this study (12 men and 7 women; age M = 20.1 years, SD = 4.02). Fifteen participants were in the first semester of their psychology studies and four students were prospective students in the last phase of pre-university education. The group of domain experts consisted of sixteen teachers (6 men and 10 women; age M = 38.2 years, SD = 11.39) from a psychology faculty of a Dutch University who volunteered to participate in this study. Six of the

teachers were from the clinical psychology research group, five were from the health psychology group and five were from the work and organisational psychology group. All were university teachers with a PhD in Psychology who had expertise on the basic topics of our search tasks. Participants were given a small financial reward of 15 Euro for their participation.

2.2 Materials

2.2.1 Questionnaire on the use and functioning of search engines. Because we intended the domain experts and domain novices to differ only in terms of their domain knowledge and not in terms of their declarative knowledge about the use of search engines, we checked for the latter using a questionnaire consisting of ten statements. For each statement the participants were instructed to indicate if it was true or false or to tick 'I do not know'. Examples of statements are: 'When searching with a search engine, the sequence of search terms does not matter', 'All search engines on the WWW function in the same manner.' The maximum score a participant could achieve was 10, and the minimum score was 0. Cronbachs' alpha of the questionnaire was .75.

2.2.2 Free recall task on topic knowledge. As a check on participants' prior knowledge regarding the two specific psychology topics addressed in the search task (i.e., the reliability of human memory and the existence of altruism), they were asked to write down in about 200 words what they knew about these topics. Performance was scored by counting the correct statements participants made. There was no maximum score defined on beforehand. The number of correct statements per participants was just counted. Two researchers scored the recall task and the inter-rater agreement was .78. They discussed the disagreements in scores. Examples of

statements mentioned by the experts when performing the memory task were: 'prof. Loftus is an expert in this field', or 'DRM-research shows evidence for false memories'.

2.2.3 Information problem-solving tasks. Participants were informed that they would have ten minutes to select 5 websites from a search engine results page that they would use in writing an argumentative article of about 1200 words for a popular psychology magazine (note that participants were only asked to select the sources, but not to really write the article). They were asked to do this for two different topics: the reliability of human memory and the existence of altruism. The task description for the memory task was: "You were asked to write a 1200 words article for a popular psychology magazine on the reliability of human memory. The question is: How reliable is human memory in a legal trail?" To answer this question the participant needs to find arguments describing why memory is or not is reliable in a legal trail. For each task a Google-like result page (SERP) with 17 or 18 links, respectively, was composed. The SERPs and all linked pages were put offline to make sure that all participants would work in exactly the same experimental setting. Both SERPs consisted of a selected mix of existing Websites. In this collection, links varied as much as possible in the quality of appearance, contents, reliability, and topicality. For example, Websites of universities with primary sources, sites of newspapers with also scientific information (secondary sources), blogs and magazines were used. The mix of these different types of sources (and as a consequence the probably perceived difficulty) was as equal as possible for both tasks. For both tasks (the task on human memory and the task on altruism) participants were asked to select from the SERP the best five websites and prioritize them. The tasks were of equal complexity according to the task complexity levels of Mosenthal (1998). Both tasks required information that can be typified as evidence; evidence or arguments to show that altruism exists and evidence to show that our

memory may be unreliable. In both tasks conflicting information could be found in the sense that there were nuance differences in the arguments. So, the conflicting information is not black and white, but as a metaphor, there are different shades of grey. Furthermore, according to Mosenthal's (1998) task classification both tasks can be characterized as an integration task, in which different arguments (or pieces of information) needed to be evaluated and combined to come to an answer. Important to mention is that complexity is not the same as difficulty. Complexity refers to the task specific characteristics and the level of complexity is in this study operationalized by using the classification of Mosenthal. The difficulty is what the reader perceives subjectively and not an objective characteristic of the task.

2.2.4 Cued retrospective reporting instruction. In CRR participants report the thoughts they had during task performance based on a replay of their task performance on which their eye-movements are overlaid (Van Gog et al. 2005). In order to provide participants with sufficient time to verbalize their thoughts, the speed of the replay was adjusted (to half of the original speed). Instructions and prompts were worded as similar as possible to the instructions in the TA condition (cf. Van Gog et al., 2005): 'This is a record of your eye movements and your actions on the PC. I am going to replay it at half the original speed, please watch it and tell me what you were thinking during task performance. Please verbalize everything you were thinking, and do not mind my presence in doing so, even when curse words come to mind for example, these should also be verbalized. Act as if you were alone, with no one listening, and just keep talking'. Whenever participants stopped verbalizing their thoughts, the experimenter prompted them after 5 seconds by saying, 'Please try to keep talking'. To familiarize participants with CRR, they were given a warm-up task: to select the best Website out of a SERP consisting 10 Websites to answer a question on an unrelated topic (traffic in The Netherlands) while their eye movements

were being recorded and then reporting verbally on their thoughts during the task while observing the replay (using the same wording in the CRR instructions)

2.2.5 Thinking aloud instruction. Participants were asked to verbalize everything they were thinking during task performance, using the following instructions: "Thinking aloud means that you should really think aloud, that is, verbalize everything that comes to mind, and not mind my presence in doing so, even when curse words come to mind for example, these should also be verbalized. Act as if you were alone, with no one listening, and just keep talking." Whenever participants stopped verbalizing their thoughts, the experimenter prompted them after 5 seconds by saying, 'Please try to keep talking'. Instructions and prompts were worded in line with the standards described by Ericsson and Simon (1993). To familiarize participants with thinking aloud, they were given a warm-up task: to select the best Website out of a SERP consisting 10 Websites to answer a question on an unrelated topic (climate in The Netherlands) while thinking aloud (using the same wording in the TA instructions).

2.2.6 Audio recording equipment. Participants' verbalizations were recorded digitally using Audacity 1.2.6 (see http://audacity.sourceforge.net) and an external microphone attached to the stimulus PC.

2.2.7 Eye tracking equipment. Eye movements were recorded with a remote 50 Hz Tobii 1750 eye tracker with infrared-cameras built into a 17-inch monitor (set to a resolution of 1024 x 768 pixels), operated by ClearView software (see www.tobii.com). The minimum fixation duration was set to 100 ms with a fixation radius of 30 pixels (cf. Cutrell & Guan, 2007; Gerjets et al., 2011). The screen capture recording mode was used, which meant that not only the eye movements, but the entire task performance process including possible mouse and keyboard operations was recorded. The eye tracking data of two experts and four novices had to be

excluded from data analyses, as they were of insufficient quality (i.e., very low tracking ratio; this left suitable eye-tracking data of n = 15 domain novices and n = 14 domain experts).

2.3 Design and procedure

The experiment had a two-factorial mixed-model design, with domain expertise as the between-subjects factor and verbal reporting technique (CRR vs. TA) within-subjects factor. The order of reporting technique was fixed, with CRR on the first task and TA on the second task, because having participants think aloud first might have undesirable carry-over effects to the second task. Moreover, because of the length of the cued reporting procedure (10 min task performance + 20 min reporting), whereas this risk was deemed much smaller with TA as this is done concurrently with the second task (i.e. finished in 10 min). The order of the two search tasks was counterbalanced between participants, though, to ensure that any differences that might be found between the reporting techniques would not be due to the specifics of the tasks (i.e., half of the participants completed the task about the reliability of human memory under CRR and the task about altruism under TA conditions; for the other half the search task order was reversed, but not the reporting technique).

Participants were tested in individual sessions of approximately 60 minutes. First, they filled out the questionnaire about knowledge of how to search on the Internet and wrote down (free recall) what they knew about the topics of the existence of altruism and reliability of human memory. Then the first task began. Participants received the CRR warming-up task with the instructions as described above. After the warming-up task the participants were given ten minutes to work on the main task (i.e., either the memory task or the altruism task). After 8 min. were up, the experimenter warned participants that there were only 2 min. left and that they might want to start on their rank-ordering of the websites. During the task participants' eye

movements and actions on the computer were recorded. Then, participants were instructed to report their thoughts while reviewing their own eye movements on half speed (i.e., 20 min.). Students were not allowed to stop the playing of the eye-movements. Research (Van Meeuwen et al.) shows that replaying the eye-movements on 75% of the normal speed gives participants enough time to elaborate their thoughts. In our study we used 50% of the normal speed to ensure participants have enough time. Subsequently, the second task began. Participants received the instructions on how to think aloud and practiced it on the TA warming-up task. Next, they were given another ten minutes to work on the second search task while thinking aloud, during which their eye-movements were also recorded. So in both tasks (CRR and TA) eye-movements were recorded for studying H2, H3 and H4. Furthermore in CRR-task the eye-movements were replayed to the participants to act as a cue for retrospective reporting. After the experiment the verbal protocols of the CRR and TA were transcribed based on the audio recordings.

2.4 Data-analysis and dependent measures

2.4.1 Eye-movement data analysis. For each search result a polygonal "area of interest" (AOI) covering the title, excerpt, and URL of the search result was defined, to determine whether or not a participant was looking at a search result. A search result was considered as visually scanned if there was at least one fixation (≥ 100 ms) within the search result. To analyze how exhaustively participants considered the different search results presented on the SERP before accessing the first Website, the *percentage of search results visually scanned before the first click*, that is, before they accessed the first Website, was determined (cf. Kammerer & Gerjets, 2014; Pan et al., 2007).

In addition, to analyze whether participants attended to source information on the Websites that they had selected as the best five, AOIs were defined on all areas on the Websites that provided source cues (i.e., the Website logo and author or publisher information). For each of the Websites a participant selected as one of the best five, it was determined whether or not the participant had paid attention to at least one source cue, to analyse the number of Websites for which source information scanned. Again, a visual inspection was defined as at least one fixation (≥ 100 ms) within a source AOI. The *percentage of Websites for which source information was scanned* was defined as dependent variable.

2.4.2 Log-file analysis. To analyze how many of the Websites provided by the search engine were accessed by the participants to solve the task, the *percentage of search results clicked on* during the 10-minute Web search was determined. In addition, for the Websites a participant selected as the best five, it was determined how much *time the participant spent on average on each site*.

2.4.3 Task performance scoring system. To indicate whether participants selected the five most appropriate Websites to accomplish the task, not only the rank of the sites was taken into account, also the trustworthiness of the sites was incorporated. Taken only the rank of the site would be less accurate, because the difference for instance between the sites on rank 1 and 2 concerning trustworthiness, is could different than that of rank 2 and 3. Taken the trustworthiness into account by given the site a trustworthiness score gives a more accurate result of participant's selection of sites. Two raters (the first and third author, both experts in information problem solving and evaluation sources and information) independently scored the trustworthiness of each site provided in the SERP on 10-point scales according to a set of evaluation criteria (e.g., the author's expertise, the type of source (primary, secondary), and the relevance and currency of

the information). Thus, each site could receive a mean trustworthiness score between 1 (very low trustworthiness) and 10 (very high trustworthiness). The correlation reflecting the agreement between the two raters was high ($r = .99$, $p < .001$). To determine the trustworthiness score of each site the scores of both raters were collapsed. The scores for the different websites ranged from 2.97 to 8.65 ($M = 5.63$, $SD = 1.63$) for the "altruism" task, and from 1.93 to 9.12 ($M = 6.16$, $SD = 2.04$) for the "reliability of human memory" task. Participants' task performance score reflecting the appropriateness of their selection was calculated based on the trustworthiness score of a selected site and the participant's ranking of this site (ranks 1 to 5). The score of the website with the highest rank got multiplied by five, the second highest ranked website by four, and so on. For example, suppose a participant selected five sites with the following trustworthiness scores: 8.26 (rank 1), 5.23 (rank 2), 6.41 (rank 3), 7.53 (rank 4), and 6.17 (rank 5). The participant performance score was calculated accordingly: (8.26*5) + (5.23*4) + (6.41*3) + (7.53*2) + (6.17*1) = .102,68 This sum score was then divided by 15 to get an average score between 0 and 10 again. For the participant the task performances score would be 6.85.

2.4.4 Coding Scheme for Evaluation of Search Results, Websites, and Information.

The coding scheme from the study by Walraven, Brand-Gruwel, and Boshuizen. (2009) was adapted to score the CRR and the TA protocols with regard to the criteria participants used for evaluation when searching information from a SERP. As main classification we distinguished between participants' evaluation of search results, the evaluation of a Website, and the evaluation of information provided in a site. Within this classification two levels of detail in participants' evaluations were distinguished: (1) superficial evaluations, for instance "Hmm, no, not this one.", and (2) specific evaluations based on a particular criterion, for instance, "not so much

information is available here, so I don't think I can use this." This latter example was rated as an evaluation of the usefulness of information based on the criterion 'amount of information'.

To validate the scoring scheme, 10% of the protocols (half CRR and half TA) were scored by two raters resulting in an interrater reliability of .76 (Cohen's Kappa). One rater coded the remaining protocols. Reliabilities were calculated on the level shown in Table 1. This is the most detailed coding scheme and coding on this level must be reliable to be able to aggregate the data used in Table 2.

3. Results

3.1 Control variables

To check that the domain experts and domain novices did not differ in their knowledge on the use and functioning of search engines, a t-test was conducted that revealed no significant differences between the two groups (experts: $M = 4.63$, $SD = 2.80$; novices: $M = 4.84$, $SD = 2.73$; $t(33) = 0.231$, $p = .818$). Furthermore, t-tests were conducted to check that the prior knowledge of the experts concerning the search task topics was indeed higher than that of the novices. This was the case both for the topic 'reliability of human memory' (correct statements: experts $M = 6.47$, $SD = 2.70$ and novices $M = 2.61$, $SD = 1.04$, $t(32) = -5.103$, $p < .001$, partial $\eta^2 = .507$); and for the topic 'existence of altruism' (correct statements: experts $M = 3.38$, $SD = 1.26$, and novices $M = 0.94$, $SD = 1.21$, $t(32) = -5.74$, $p < .001$, partial $\eta^2 = .503$). Thus, it can be concluded that as we expected, experts and novices indeed differed in their domain expertise central to the tasks of this study, but not in their knowledge on the functioning of search engines. Furthermore, the division between male and female was not equal in the groups (experts and novices). We analysed if gender differences on the control variables occurred. This was not the case. All p-

values of the interaction effects 'expertise x gender' in the analyses of variance are respectively .99 for search engine knowledge, .73 for prior knowledge on the reliability of human memory and .30 for prior knowledge on the topic altruism.

3.2 Evaluation criteria used by domain novices and domain experts during CRR and TA

Table 1 gives an overview (means and standard deviations) of the criteria used by the experts and the novices under the different reporting techniques (CRR and TA). For the analyses, data across the two task scenarios ('altruism' and 'memory') were collapsed for each reporting technique. The criteria are categorized by the evaluation of the usefulness and reliability of (1) the search results, (2) the Websites, and (3) the information found in a site. Furthermore, as mentioned in section 2.4.4 the evaluations were scored on two levels, that is, superficial and specific. At the top of the table also the total amount of evaluation statements made by the experts and novices in the different reporting conditions are presented.

INSERT TABLE 1

As can be seen from Table 1, overall more utterances about evaluation of sources and information were expressed in the CRR condition than in the TA condition. Most of the evaluations when judging the search results, Websites, and information concerned the expected connection to the task of writing the article for the popular psychology magazine. When looking further into how experts and novices in a domain evaluate the *usefulness* of the search results in a SERP it can be seen that not many other criteria are used than the connection to the task. Concerning the *reliability* it can be seen that people evaluate the search results in a SERP and the Websites especially based on the kind of source. For instance whether a Website is from a

university, a blog, or a newspaper. Overall, reliability is not often questioned. The Table shows that the means are very low (often below 1) and many participants not even used certain criteria.

When aggregating the judgments concerning reliability, usefulness and technical issues, respectively, across search results, Websites, and information, the differences between domain experts and novices can be calculated on a specific level, that is the use of specific criteria when evaluating. Data is aggregated because testing multiple hypotheses and running analyses on all criteria would lead to incorrect interpretations. Furthermore, the fact that some criteria were not or hardly used also would lead to incorrect results. Table 2 gives an overview of the means, standard deviations, minimum and maximum for the specific level on the used criteria for usefulness and reliability and also includes the amount of superficial evaluations.

INSERT TABLE 2

For the variable *total superficial evaluations* a mixed-model ANOVA with domain expertise (experts vs. novices) as between-subject factor and reporting technique (TA vs. CRR) as within-subject factor showed a significant main effect for domain expertise, $F(1,32) = 8.86$, $p = .006$, partial $\eta^2 = .22$. The novices expressed more superficial evaluations. Also a main effect for the factor reporting technique was found, $F(1, 32) = 10.29$, $p = .003$, partial $\eta^2 = .24$. The CRR technique elicited more such evaluations. No interaction was found, indicating that the two techniques did not show different effects depending on the expertise level of the participants.

For the variable *total specific evaluations on usefulness* the analysis revealed only a main effect for reporting technique, $F(1, 32) = 40.52$, $p < .001$, partial $\eta^2 = .56$. For both levels of expertise the CRR condition elicited more specific evaluations on the usefulness of information than TA. No main effect for domain expertise and no interaction effect were found. The lack of significant interaction effect may be due to the lack of power. Examining Table 2 it can be seen

that for this variable the TA method is in favour of the novices. In the TA condition the novices scored higher than the experts and in the CRR the experts scored higher than the novices.

Finally, for the variable *total specific evaluations on reliability* the analysis showed a significant main effect for domain expertise, $F(1, 32) = 5.10$, $p = .031$, partial $\eta^2 = .14$. The experts expressed more specific evaluations on the reliability of search results Websites and information than the novices. Also a main effect for the factor reporting technique was found, $F(1, 32) = 7.85$, $p = .009$, partial $\eta^2 = .20$. The CRR technique elicited more evaluations than TA. No interaction was found indicating that the two techniques did not have different effects depending on the expertise level of the participants.

3.3 Evaluation behavior of domain experts and novices as measured through eye-tracking and log-file data

Table 3 shows (besides the task performance scores) means and standard deviations of participants' evaluation behavior during task processing as measured through eye-tracking and log-file data as a function of domain expertise and reporting technique.

INSERT TABLE 3

During the 10-minute Web search participants on average visited 60.54% ($SD = 16.56$) of the Websites provided in the SERPs, without any differences found for domain expertise or reporting technique, nor a significant interaction between the two factors (all Fs < 1).

With regard to the *percentage of search results visually scanned before the first search result was clicked* we found significant differences between experts and novices, $F(1, 27) = 15.81$, $p < .001$, partial $\eta^2 = .37$. Domain experts visually scanned a significantly higher

percentage of search results on the SERP before accessing the first Website (M = 73.82%, SD = 32.53) than novices (M = 34.86%, SD = 31.12), indicating a higher scrutiny of the experts in evaluating the search results. Besides, there was no main effect of reporting technique, $F(1, 27) = 2.24$, $p = .146$, nor a significant interaction between expertise and reporting technique, $F < 1$.

When analyzing participants' *task processing of selecting the five best sites, with regard to the percentage of Websites for which source information (e.g., the logo or author information) was scanned*, no differences were found between domain experts and novices, $F < 1$. However, there was a significant main effect of reporting technique, $F(1, 27) = 4.40$, $p = .046$, partial $\eta^2 = .14$, with participants in the TA condition attending to source information in a significantly higher percentage of the sites (M = 79.60%, SD = 20.31) than in the CRR condition (M = 66.72%, SD = 26.10). There was no support for an interaction between domain expertise and reporting technique, $F < 1$.

With regard to the *average time spent on a selected Website*, the ANOVA showed a significant main effect of domain expertise, $F(1, 33) = 4.47$, $p = .042$, partial $\eta^2 = .12$. Domain experts spent less time (M = 47.27 s, SD = 23.19), that is, they decided to select a site faster than novices (M = 69.07 s, SD = 36.80). Besides, there was no effect of reporting technique, $F < 1$, nor a significant interaction between domain expertise and reporting technique, $F(1, 33) = 1.45$, $p = .237$.

3.4 Task performance of domain experts and novices and relation with use of evaluation criteria

Participants' task performance, that is, the appropriateness of their selection of the five best Websites, was calculated based on the trustworthiness score of a selected site and the

participant's ranking of this site (ranks 1 to 5). Table 3 also presents the task performance scores of experts and novices for the CRR and the concurrent TA-tasks.

The ANOVA showed a significant main effect of domain expertise, $F(1, 33) = 4.01$, $p = .05$, partial $\eta^2 = .11$. As expected, domain experts yielded significantly higher performance scores ($M = 7.03$, $SD = 0.77$) than domain novices ($M = 6.56$, $SD = 0.97$). Besides, there was no effect of reporting technique nor a significant interaction between domain expertise and reporting technique, Fs < 1.

Furthermore, a significant correlation was found between the *performance score* and *total specific evaluations on reliability,* $r =. 293$, $p = .044$. Furthermore, a negative correlation was found between *performance score* and *total superficial evaluations* $r =. -303$, $p = .036$. In contrast, no significant correlations were found between the performance and specific evaluations of the usefulness. Looking at these correlations per condition it can be seen that for the novices no significant correlation could be found. For the expert a significant negative correlation between *performance score* and *total specific evaluations on usefulness* was found, $r = -.531$, $p = 017$.

To get more insight in the websites experts and novices selected, a Mann-Whitney U test was conducted. The sites participants put on the first place in their ranking were analysed. The sites in the ranking between the experts and novices were compared. This was also done for the websites the participants put on the second to fifth place. The Mann-Whitney U test was conducted for both tasks (Altruism and Reliability of human memory).

For the Altruism task it was found that the experts (Mdn = 2,5) put more trustworthy sites than the novices (Mdn = 9.0) on the first place of their ranking ($U = 65.5$, $p = .012$). Looking into the sites selected by the groups it seems that 44% of the experts selected the two primary

scientific sources of top researchers and put them on top of their ranking, while only 16% of the novices selected these sites. Furthermore, 53% of the novices prioritized a personal site of an unknown philosopher. This site addressed the topic, but there were no references to research and the authors' background was not mentioned. Only 19% of the experts selected this site.

For the Reliability of human memory task the same pattern was found, although the significance was at a 10% level and thus a trend. The experts (Mdn = 1.5) put more trustworthy sites than the novices (Mdn = 4.0) on the first place of their ranking ($U = 87.5$, $p = .098$). Looking into the sites selected by the groups it seems that 57% of the experts selected the two primary scientific sources of top researchers and put them on top of their ranking, while only 21% of the novices selected these sites. The site the novices put on first place was a mix of different sites but all were less trustworthy.

4. Discussion

4.1 Evaluation behavior of domain experts and domain novices

This study aimed to gain insight in the impact of domain knowledge on individuals' evaluation behavior when searching the Web to prepare a magazine article, a scholastic task. Furthermore the study focused on different techniques to measure the underlying processes of the evaluation of sources. Participants engaged in one search task while thinking aloud as well as in another one in which the method of cued retrospective reporting was used after task processing.

From the verbal utterances gained in this study, first of all it can be concluded that by far the most evaluation utterances of the participants concerned the connection to the task. This is in line with the theory of relevance in the text processing (McCrudden & Schraw, 2007) and with research of Braasch, Lawless, Goldman, Manning, Gomez, and MacLeod, (2009). From a

cognitive perspective information processors need to search for the most relevant information to achieve optimal cognitive efficiency. Optimal efficiency is of importance because nowadays people encounter massive amounts of information, yet have extremely limited cognitive resources to process and remember this information (Ericsson & Kintsch, 1995). One way to adapt is to grade information for the amount of contextual relevance it contains and to prioritize high-relevance information. This is what people do when evaluating the relevance and connection to the task.

In contrast, participants hardly evaluated the reliability of sources and when doing so they mainly questioned the kind of source (e.g. a newspaper, a blog, a forum, etc.). This may be due to the fact that these different Websites are designed in certain ways and that catches the eye. These findings are in line with previous research with university or secondary-school students (e.g., Gerjets et al., 2011; Walraven, Brand-Gruwel, & Boshuizen, 2009).

Concerning the hypotheses it was expected that domain experts would express less superficial evaluations (H1) and more specific utterances concerning reliability evaluations (H2) than domain novices. Both hypotheses were confirmed. Domain novices expressed more utterances on a superficial level, mentioning more evaluations like "seems nice", "no, I don't like this one" and domain experts expressed more specific utterances to judge the reliability of sources, such as “I know this author, he is an expert in the field" and “This is a site of a university, seems reliable”. These findings are in accordance with findings of previous studies by Rouet et al. (1997), Stadtler, et al. (2013), Stanford et al. (2002), and Wineburg (1991). Also, as stated in the introduction, the study of Salmerón et al. (2013) found that students with higher domain knowledge made decisions about which Websites to use for further study by bookmarking them, based on trustworthiness of the Websites more than students with little

domain knowledge. They also less often based their decision on the link position in the search engine results page (SERP). Furthermore, experts tend to bookmark first a primary source (e.g. scientific journal), whereas novices tended to pick a blog post made of different (potentially primary) sources. This is in line with research of Von der Mühlen, Richter, Schmid, Schmidt, and Berthold. (in press). In their research they examine how students of psychology and scientists evaluate arguments, while think-aloud. Results indicate that students, compared to scientists, have difficulties with evaluating and base them on intuition and opinion rather than the internal consistency of arguments. In our study we used science topics, but the study of for instance Wineburg (1991) used history topics. However Goldman (2012) stated that different disciplines have differentiated literacy practices. She argues that reading to learn in the different disciples (science, history and literature) also the evaluation of information (on a more detailed level) differ. For instance evaluation in literature involves evaluation of poem, stories and drama, while evaluation in history and science involves judging multiple sources of information presented in diverse formats and media. In future research differences between domains or disciples should be studied to gain more insight in different evaluation criteria and specific instructional strategies.

Concerning the eye-tracking and log-file data it was expected that domain experts have more search results fixated before the first Website being accessed from the SERP (H3), are more likely to attend to source information to decide that a Website should be selected (H4), and spend less time on a Website to decide that it should be selected then domain novices (H5). Hypotheses 3 and 5 were confirmed. According to Kammerer and Gerjets (2014) a higher number of search results fixated before the first click is an indication for increased evaluations of the search results. Moreover, the results regarding H5 are in line with earlier research by

Hölscher and Stube (2000). They also found that domain experts spent less time on Websites and argued that domain experts do need less time than domain novices for reading and to decide whether a site is appropriate or not. However, in contrast to our expectations in H4, domain experts weren't more likely to attend to source information in the Websites that they selected as one of the best five than domain novices. Still, as reported above experts mentioned more specific evaluations on reliability.

Concerning the task performance it was expected that domain experts would select more trustworthy Websites than domain novices (H6). Although the difference in the means was not large, it was significant. The hypothesis was confirmed. Further analyses also revealed a correlation between the performance score and the amount of specific evaluation concerning the reliability of sources and information. The more the participants judge the reliability using specific criteria, the higher the performance score. Furthermore, the variable 'superficial evaluations' was negatively related to task performance. For experts it can be said that task performance had a negative relation with the amount of evaluations on usefulness. When taking a closer look at the selected Websites it can be concluded that the experts prioritized the scientific primary sources more often than the novices, although they did not attend more to source information then the novices. Overall it can be concluded that domain expertise has an impact on individuals' evaluation behavior during Web search, such that domain expertise is associated with a more sophisticated use of evaluation criteria to judge the reliability of sources and information and leads to better selection of reliable information. Furthermore, the more use of specific criteria and less use of superficial criteria to judge the trustworthiness, the better the performance. This finding is in line with research of Goldman et al. (2012) and Wiley et al. (2009) who also found a relation between source evaluation behavior and learning outcomes.

4.2 Thinking aloud and Cued Retrospective Reporting compared

In this study two verbal reporting techniques to uncover the processes of evaluating Web information were compared. CRR and TA were used to explore if both techniques lead to the same conclusions concerning differences between domain experts' and novices' evaluation behavior. The overall amount of utterances during CRR was higher than during TA; for all three types of utterances analyzed, namely total superficial evaluations, total specific evaluations on usefulness, and total specific evaluations on reliability more utterances were expressed in the CRR condition. Potentially, this is a consequence of the fact that more time was available, as in the CRR condition participants expressed their thoughts while looking at the record of their performance process and eye movements at half speed. The fact that in the CRR condition more thoughts were expressed than in the TA condition differs from the findings by Van Gog et al. (2005). However, in contrast to the study by Van Gog et al., participants in our study were given twice as much time to report in the CRR condition than in the TA condition (because the replay was slowed to half speed to give them sufficient time; in the Van Gog et al. study the replay was in real time). Another explanation can be found using the Memory Subsystems as Processes (MEM) framework by Johnson and colleagues (Higgins & Johnson, 2012; Johnson & Hirst,1993) on source memory. They claim that readers may reconstruct sources at the time of retrieval using different memory traces of the information processed, and not only at the time of encoding. Under this framework, sourcing may depend on the time available to perceive and reflect on the presented information. Participants in the CRR had the chance to reconstruct source memory at the time of retrieval (in twice as much time), and this may have increased their evaluations on reliability. Participants in the TA condition, by contrast, may have focused on

semantic information at the time of encoding, since at the time of encoding semantic and more reflective processes may compete for memory resources. Since they did not have the chance to rely on memory retrieval to reconstruct source information, they may have produced less evaluations on reliability at the time of encoding. Finally, this finding might simply result from the fact that CRR was engaged in first, and TA second (we will return to this below, under limitations).

Nevertheless, this is interesting for researchers who intend to uncover process information. Although more utterances were produced in the CRR condition, no interaction effects were found between the used technique and the level of expertise. However it should be mentioned that although no significant interactions were found the data show some indications that there may be differences in the reporting techniques as a function of level of expertise, and perhaps this study lacked sufficient power to detect significant interaction effects. Thus, although with caution, it can be concluded that the used techniques do not lead to different findings for experts and novices in terms of the elicited utterances.

Furthermore, the verbal reporting techniques did not seem to affect participants' task performance. It has been suggested that thinking aloud while performing a task would not alter the process but might slow it down or affect the eye movements made (Holmqvist et al., 2011). However, as indicated by the eye-tracking and log-file data, during task processing the verbal reporting techniques did not affect the number of search results fixated before the first click or the time spent on the Websites that participants selected as one of the best five. Only with regard to the percentage of these Websites for which source information (e.g., the logo or author information) was scanned, a significant difference was found between the two verbal reporting techniques. In the TA condition for a higher percentage of the selected Websites source

information was attended than in the CRR condition. This possibly indicates a somewhat more conscious or controlled evaluation behavior elicited by the TA procedure. However, this effect was not moderated by domain expertise.

Although further research is needed, the results of the present study do give a first indication that the use of CRR or TA do not lead to differential effects depending on the level of domain expertise, and, do not result in differential conclusions for the different levels of domain expertise. Decisions about which technique to use, depends on the research questions at stake. When you want to gain insight in cognitive or metacognitive processes during task performance, one of the first issues that arise is if there is a visual component of importance in the task. In CRR participants express their thoughts looking at their eye-movements, so these eye-movements should have an added valued. When performing web search tasks eye movement data, and the use of the eye movements as cues in CRR is very useful. Furthermore, it is of importance to take into account the length of the task. Longer tasks are not advisable for using CRR, because remembering the cognitive processes while looking at the eye-movement cues will become harder over time. On the other hand, when eye-movement data is an important data source for answering the research questions, then one should be aware that thinking aloud during task performance can have an impact on the data, as in our case on the attention to source information.

4.3 Limitations and further research

Some limitations of the present research should be addressed. First, the number of participants was limited and the experts were teachers, who in general might be more critical than students irrespective of their domain knowledge. This may as a consequently cause that

teachers are also more skilled searchers. In previous works Rouet makes the claim that experts not only differ on knowledge on the topic, but also on procedural knowledge about proper skills for document use in a given field (Rouet et al., 1997). In our study we only measured declarative knowledge, why procedural knowledge maybe could have given more can insight in people's use of the skills. Hence, the age and profession of the participants may have affected the results. Related to these issues, it can be questioned whether it is only the domain experts' higher content knowledge that caused the effects or whether it also their higher source knowledge (e.g., knowledge about how to judge the trustworthiness of sources in general). Furthermore, the experts might have a better idea about what it means to write an article for a popular magazine, which could have also influenced their search behaviour. In future research experts and novices could be recruited from the same kind of population. It is an idea to recruit university students from different fields of expertise in their final year. Psychology and Physics students would be asked to accomplish a psychological oriented and a physics oriented task and serve as expert and novice as well.

Second, in the two groups (experts and novices) the division of male and female was not equal. We checked for gender differences on the control variables and no differences were found. However, one could question if results could have been influenced by these gender differences in the two groups. Our research questions did not focus on the differences in gender. Earlier research of Walhout, Brand-Gruwel, Van Dijk, Jarodzka, De Groot, and Kirschner, (2015) studied the effect of gender in web searching behavior of tenth grade students and found no differences.

Third, in the design of the experiment the tasks (altruism and reliability of human memory) were counterbalanced, but the verbal reporting techniques (CRR and TA) were not. As addressed

above, the idea was that engaging in TA first could have undesirable carry-over effects to the second task that had to be performed in silence, and could therefore affect the CRR data (i.e., turn them into a report of previous mental verbalizations instead of thoughts as they would normally occur during task performance). In a follow-up study, however, it would be wise to also counterbalance these conditions (potentially running them on different days so as to minimize carry-over effects) and furthermore pay attention to the length of the task, because think aloud using a short 10 minutes task could have led to more cognitive load in this condition. This could also rule out the possibility that the finding that CRR led to more data than TA would merely be a consequence of participants getting tired as the experiment progresses, which could have been the case in the present study. If this holds true even after counterbalancing the methods, then CRR could be a valuable tool for research in this field.

5. References

Braasch, J. L. G., Bråten, I., Strømsø, H. I., Anmarkrud, Ø., & Ferguson, L. E. (2013). Promoting secondary school students' evaluation of source features of multiple documents. *Contemporary Educational Psychology*. 38, 180–195. doi:10.1016/j.cedpsych.2013.03.003

Braasch, J. L. G., Lawless, K. A., Goldman, S. R., Manning, F. H., Gomez, K. W., & MacLeod, S. M. (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., Vibert, N., & Britt, M. (2012). Readers' use of source information in text comprehension. *Memory & Cognition, 40*, 450-465. doi:10.3758/s13421-011- 0160-6.

Brand-Gruwel, S., & Stadtler, M. (2011). Solving information-based problems: Searching, selecting and evaluating information. *Learning and Instruction, 21*, 175-179.

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving by experts and novices: analysis of a complex cognitive skill. *Computers in Human Behaviour, 21*, 487-508.

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

Conrad, F. G., Blair, J., & Tracy, E. (1999). Verbal reports are data! A theoretical approach to cognitive interviews. In *Proceedings of the Federal Committee on Statistical Methodology Research Conference*, Tuesday B Sessions, Arlington, VA, 11–20.

Cutrell, E., Guan, Z. (2007). *What are you looking for? An eye-tracking study of information usage in Web search.* In The Proceedings of ACM CHI'07, New York: ACM Press. 407 - 416.-1061.

Ericsson, K. A., & Kintsch, W. (1995). Long-term working memory. *Psychological Review, 102(2)*, 211–245

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

Gerjets, P., Kammerer, Y., & Werner, B. (2011). Measuring spontaneous and instructed evaluation processes during web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction, 21*, 220-231.

Goldman, S. R. (2012). Adolescent literacy: Learning and understanding content. *Future of*

Children, 22, 89–116.

Goldman, S. R., Braasch, J. L. G., Wiley, J., Graesser, A. C., & Brodowinska, K. (2012). Comprehending and Learning From Internet Sources: Processing Patterns of Better and Poorer Learners. *Reading Research Quarterly, 47*(4), 356–381. doi: 10.1002/RRQ.027

Higgins, J.A., & Johnson, M.K. (2012). Some thoughts on the interaction between perception and reflection. In J.M. Wolfe & L. Robertson (Eds.), *From perception to consciousness: Searching with Anne Treisman* (pp. 390-397). New York: Oxford University Press.

Holmqvist, K., Nyström, N., Andersson, R., Dewhurst, R., Jarodzka, H., & Van de Weijer, J. (2011). Eye tracking: a comprehensive guide to methods and measures, Oxford, UK: Oxford University Press.

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

Jamali, H. R., & Asadi, S. (2010). Google and the scholar: the role of Google in scientists' information-seeking behaviour. *Online information review, 34*, 282-294. Baslev Baslev

Johnson, M.K., & Hirst, W. (1993). MEM: Memory subsystems as processes. In A.F. Collins, S.E. Gathercole, M.A. Conway, & P.E. Morris (Eds.), *Theories of memory* (pp. 241-286). Sussex, UK: Erlbaum.

Kammerer, Y., & Gerjets, P. (2012). Effects of search interface and internet-specific epistemic beliefs on source evaluations during web search for medical information: An eye-tracking study. *Behaviour & Information Technology, 31,* 83-97.

Kammerer, Y., & Gerjets, P. (2013). The role of thinking-aloud instructions and prior domain knowledge in information processing and source evaluation during Web search. In M. Knauff, M. Pauen, N. Sebanz, & I. Wachsmuth (Eds.), *Proceedings of the 35th Annual*

Conference of the Cognitive Science Society (pp. 716-721). Austin, TX: Cognitive Science Society.

Kammerer, Y., & Gerjets, P. (2014). The role of search result position and source trustworthiness in the selection of Web search results when using a list or a grid interface. *International Journal of Human-Computer-Interaction, 30,* 177-191.

Kobayashi, K. (2009). The influence of topic knowledge, external strategy use, and college experience on students' comprehension of controversial texts. *Learning and Individual Differences, 19*, 130-134.

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

Kuusela, H., & Paul, P. (2000). A comparison of concurrent and retrospective verbal protocol analysis. *The American Journal of Psychology, 113*, 387–404.

Lawless, K. A., Goldman, S. R., Gomez, K., Manning, F., & Braasch, J. (2012). Assessing multiple source comprehension through Evidence Centered Design. In J. P. Sabatini, T. O'Reilly & E. R. Albro (Eds.), *Reaching an understanding: Innovations in how we view reading assessment (pp. 3 – 18).* Lanham, MD: R & L Education.

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

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

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.

McCrudden, M. T., & Schraw, G. (2007). Relevance and goal-focusing in text processing. *Educational Psychology Review, 19*(2), 113-139.

Metzger, M. J., Flanagin, A. J., & Zwarun, L. (2003). College student Web use, perceptions of information credibility, and verification behavior. *Computers & Education, 41*(3), 271-290.

Mosenthal, P. B. (1998). Defining prose task characteristics for use in computer-adaptive testing and instruction. *American Educational Research Journal, 35*, 269-307

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.

Purcell, K., Heaps, A., Buchanan, J. & Friedrich, L. (2013). *How teachers are using technology at home and in their classrooms.* Washington, DC : Pew Research Center's Internet & American Life Project.

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., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology, 88,* 478–493. doi:10.1037/0022-0663.88.3.478

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., Kammerer, Y., & García-Carrión, P. (2013). Searching the Web for conflicting topics: page and user factors. *Computers in Human Behavior, 29*, 2161–2171.

Schwonke, R., Berthold, K., & Renkl, R. (2009). How multiple external representations are used and how they can be made more useful. *Applied Cognitive Psychology, 23,* 1227-1243.

Stadtler, M., Scharrer, L., Brummernhenrich, B., & Bromme, R. (2013). Dealing with Uncertainty: Readers' memory for and use of conflicting information from science texts as function of presentation format and source expertise. *Cognition and Instruction, 31,* 130-150. doi:10.1080/07370008.2013.769996

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.

Strømsø, H. I., Bråten, I., Britt, M. A., & Ferguson, L. (2013). Spontaneous sourcing among students reading multiple documents. *Cognition and Instruction, 31*, 176-203. doi:10.1080/07370008.2013.769994

Tabatabai, D., & More, B. M. (2005) How experts and novices search the Web. *Library & Information Science Research, 27*(2), 222 – 248.

Taylor, K.L., & Dionne, J. (2000). Accessing problem-solving strategy knowledge: The complementary use of concurrent verbal protocols and retrospective debriefing. *Journal of Educational Psychology, 92*(3), 413-25.

Van Gog, T. (2006). *Uncovering the problem-solving process to design effective worked examples*. Doctoral Dissertation, Open University of The Netherlands, Heerlen, The Netherlands.

Van Gog, T., Kester, L., Nievelstein, F., Giesbers, B., & Paas, F. (2009). Uncovering cognitive processes: Different techniques that can contribute to cognitive load research and instruction. *Computers in Human Behavior, 25*, 325-331.

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.

Van Someren, M. W., Barnard, Y. F., & Sandberg, J. A. C. (1994). *The think aloud method: A practical guide to modeling cognitive processes.* London: Academic Press.

Van Strien, J., Brand-Gruwel, S., & Boshuizen, H. P. A. (2014). Dealing With Conflicting Information From Multiple Nonlinear Texts: Effects of Prior Attitudes. *Computer in Human Behavior*, *32*, 101-111.

Von der Mühlen, S., Richter, T., Schmid, S., Schmidt, E.M., & Berthold, K. (2016). Judging the plausibility of arguments in scientific texts: a student' scientist comparison. *Thinking and Reasoning, 22,* 221-249.

Walhout, J., Brand-Gruwel, S., Van Dijk, M., Jarodzka, H., De Groot, R., & Kirschner, P. A. (2015) Navigating through hypertext: Do sex and type of navigational support matter? *Computers in Human Behavior,46,* 218-227.

Wilson, T. D. (1999). Models in information behaviour research. *Journal of Documentation, 55*(3), 249–270.

Wiley, J., Goldman, S. R., Graesser, A. C., Sanchez, C. A., Ash, I., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in Internet science inquiry tasks. *American Educational Research Journal, 46*, 1060–1160. doi:10.3102/0002831209333183.

Wineburg, 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,* 73-87.

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

Wyatt, D., Pressley, M., El-Dinary, P. B., Stein, S., Evans, P. & Brown, R. (1993). Comprehension strategies, worth and credibility monitoring, and evaluations: cold and hot cognition when experts read professional articles that are important to them. *Learning and individual differences, 5*, 49-72.

Table 1

Mean Amount of Evaluation Utterances, Evaluation of the Usefulness and Reliability of Search Results, the Websites, and the Information provided in a Site as a Function of Expertise and Verbal Reporting Technique

	Thinking Aloud		Cued Retrosp. Reporting	
	Domain experts	Domain novices	Domain experts	Domain novices
AMOUNT OF EVALUATION UTTERANCES	26.07 (7.03)	25.11 (6.68)	38.56 (8.21)	37.63(11.23)
SEARCH RESULTS (superficial)	0.93 (1.10)	2.32 (1.70)	2.25 (2.21)	2.84 (3.08)
Usefulness (specific)				
1. Expected connection with the task	0.80 (1.01)	2.63 (1.92)	4.81 (7.04)	2.74 (2.40)
2. Comprehensibility	0.07 (0.26)	0.05 (0.23)	0.06 (0.25)	0.21 (0.42)
3. Usage of language (popular, scientific)	0.07 (0.26)	0.16 (0.38)	0.00 (0.00)	0.21 (0.42)
4. Language (English, French etc.)	0.07 (0.26)	0.11 (0.32)	0.25 (0.45)	0.21 (0.42)
5. Amount of information	0.00 (0.00)	0.00 (0.00)	0.06 (0.25)	0.00 (0.00)
Reliability (specific)				
1. Reputation	0.87 (0.92)	0.58 (0.84)	1.38 (2.73)	0.74 (0.93)
2. Format of the source (Website / pdf / ptt etc.)	0.33 (0.72)	0.26 (0.73)	0.31 (0.60)	0.21 (0.42)
3. Kind of source: Primary/secondary/tertiary	1.87 (2.07)	0.84 (1.12)	2.13 (2.28)	2.00 (1.92)
4. Up-to-dateness	0.27 (0.60)	0.05 (0.23)	0.13 (0.34)	0.11 (0.46)
5. URL	0.33 (0.49)	0.68 (0.95)	0.81 (1.22)	0.89 (1.20)
6. Familiarity	1.47 (1.64)	0.53 (0.84)	1.13 (1.31)	0.47 (0.77)
Technical (specific)				
1. Position in hit list	0.07 (0.26)	0.21 (0.54)	0.06 (0.25)	0.58 (0.96)
WEBSITES (superficial)	0.80 (1.01)	1.95 (2.01)	0.94 (1.12)	2.68 (3.15)
Usefulness (specific)				
1. Language	0.20 (0.41)	0.21 (0.63)	0.19 (0.40)	0.37 (0.68)
2. Structure	0.00 (0.00)	0.00 (0.00)	0.06 (0.25)	0.05 (0.23)
Reliability (specific)				
1. Reputation	1.60 (2.13)	0.58 (0.77)	1.56 (1.50)	0.63 (0.96)
2. Format of source (Website / pdf / ptt etc.)	0.33 (0.62)	0.05 (0.23)	0.25 (0.58)	0.26 (0.45)
3. Kind of source: Primary/secondary/tertiary	2.60 (1.45)	1.89 (1.97)	2.00 (1.37)	1.89 (1.70)
4. Up-to-dateness	0.67 (1.35)	0.21 (0.54)	1.00 (1.55)	0.68 (1.06)
5. URL	0.00 (0.00)	0.26 (0.81)	0.19 (0.40)	0.42 (0.69)
6. Familiarity	0.73 (1.03	0.32 (0.95)	0.88 (0.96)	0.16 (0.38)
Technical (specific)				
1. Appearance	0.00 (0.00)	0.42 (0.61)	0.06 (0.25)	0.74 (1.33)
2. Loading speed of Website	0.20 (0.41)	0.00 (0.00)	0.31 (0.48)	0.16 (0.38)
INFORMATION (superficial)	0.67 (0.82)	2.26 (2.38)	1.56 (.15)	3.00 (2.91)
Usefulness (specific)				

1. Expected connection with the task	4.33 (1.99)	4.53 (2.46)	8.50 (2.71)	8.89 (4.92)
2. Comprehensibility	0.20 (0.41)	0.58 (1.02)	0.94 (1.06)	1.11 (1.05)
3. Usage of language	0.13 (0.35)	0.11 (0.32)	0.25 (0.45)	0.53 (0.23)
4. Structure of information	0.13 (0.52)	0.37 (0.83)	0.38 (0.50)	0.42 (0.61)
5. Amount of information	1.33 (1.18)	1.53 (1.90)	1.88 (1.50)	1.58 (1.71)
Reliability (specific)				
1. References	1.20 (1.74)	0.32 (0.58)	0.94 (1.06)	0.95 (1.03)
2. Information is found on more sites	0.20 (0.56)	0.16 (0.38)	0.19 (0.40)	0.42 (0.69)
3. Information connected to prior knowledge	0.33 (0.82)	0.53 (0.91)	0.81 (1.11)	0.53 (0.84)
4. Goal of the information	0.00 (0.00)	0.05 (0.23)	0.13 (0.34)	0.05 (0.09)

Table 2

Superficial Evaluations, and Evaluation of the Usefulness and Reliability on a Global and Specific Level as a Function of Expertise and Verbal Reporting Technique (mean, SD, min., max.)

	Thinking Aloud		Cued Retrosp. Reporting	
	Domain Experts	Domain novices	Domain experts	Domain novices
	Mean(SD)(min-max)	Mean(SD)(min-max)	Mean(SD)(min-max)	Mean(SD)(min-max)
Total superficial evaluations	2.40 (1.96) (0-6)	6.53 (4.56) (2-17)	4.75 (3.4.7) (0-12)	8.53 (6.27) (1-25)
Total specific evaluations usefulness	7.33 (2.58) (3-12)	10.26 (3.83) (5-16)	17.38 (7.43) (4-40)	15.84 (6.69) (8-34)
Total specific evaluations reliability	14.13 (5.89) (5-25)	8.84 (5.75) (3-24)	15.69 (7.20) (3-31)	12.00 (6.52) (3-27)

Table 3

Means (and Standard Deviations) of Eye-tracking Variable, Logfile Variables and Task Performance Score as a Function of Domain Expertise and Verbal Reporting Technique

	Thinking Aloud		Cued Retrospective Reporting	
Variables	Domain experts	Domain novices	Domain experts	Domain novices
% Websites visited	58.44 (16.50)	63.40 (17.28)	59.72 (17.42)	60.15 (15.69)
% search results visually scanned before first click[1]	70.00 (37.16)	28.61 (32.42)	77.64 (27.89)	41.11 (29.81)
% of Websites for which source information scanned[1]	77.86 (20.07)	81.22 (21.10)	66.64 (31.52)	66.80 (20.97)
Average time spent on a selected Website	46.83 (22.28)	70.78 (37.56)	50.96 (22.14)	61.27 (29.67)
Task performance score	6.95 (0.91)	6.55 (1.00)	7.10 (0.63)	6.57 (0.94)

Note. [1] for the eye-tracking variables $n = 14$ experts and $n = 15$ novices were included.