How students evaluate sources and information when searching the World Wide Web for information

Open Universiteit
www.ou.nl

Abstract

The World Wide Web (WWW) has become the biggest information source for students while
solving information problems for school projects. Since anyone can post anything on the
WWW, information is often unreliable or incomplete, and it is important to evaluate sources
and information before using them. Earlier research has shown that students have difficulties
with evaluating sources and information. This study investigates the criteria secondary
educational students use while searching the Web for information. 23 students solved two
information problems while thinking aloud. After completing the tasks they were interviewed
in groups on their use of criteria. Results show that students do not evaluate results, source
and information very often. The criteria students mention when asked which criteria are
important for evaluating information are not always the same criteria they mention while
solving the information problems. They mentioned more criteria but also admitted not always
using these criteria while searching the Web.

How students evaluate sources and information when searching the World Wide Web for information.

In the last decade, the Dutch secondary educational system has undergone major changes. Instead of a system aimed at the reproduction of knowledge, the new learning environment aims at learning outcomes that should be durable, flexible, functional, meaningful and applicable. Active pedagogical methods, in which students learn by doing instead of listening, and teachers have a guiding role, fit this new learning (Simons, Van der Linden & Duffy, 2000). The teacher no longer provides students with information by standing in front of the classroom lecturing, but more and more, students actively and independently have to construct knowledge, collect information, judge information and integrate this with prior knowledge. This educational approach requires that students have a different type of skills than in traditional education. One of these skills is problem solving: a complex higher-order cognitive process that requires the modulation and control of more routine or fundamental skills (Goldstein & Levin, 1987). These routine or fundamental skills are clarifying the description of the problem, analyzing causes, identifying alternatives, assessing each alternative, choosing one, implementing it, and evaluating whether the problem was solved or not. Problem solving, a form of active participation in learning, is more satisfying than passive transfer of information from the teacher to the student and leads to enhanced retention and recall (Bransford, Brown, Cocking, 2000).Problem solving skills call upon regulation- and planning skills, also known as metacognitive skills (Perkins & Salomon, 1989; Van Merriënboer, 1997;1999).

A specific type of problem that can be distinguished is an information-based problem. Solving these kinds of problems is called the process of 'Information Problem Solving' (IPS). Being able to solve information problems means that students can identify information needs, locate

information sources, extract and organise information from each source, and synthesise information from a variety of sources (Brand-Gruwel, Wopereis & Vermetten, 2005; Eisenberg & Berkowitz, 1990; Moore, 1995). Within this IPS process evaluation of sources and information is an important sub skill, especially since students often search on the Internet for information while accomplishing learning tasks. The Web is easily accessible and students are seduced to cut and paste the information without evaluating it (Grimes & Boening, 2001; Rothenberg, 1998), resulting in reports that lack quality. A lack of evaluation of results, information and sources can affect students' learning and schoolwork.

The aim of this study is to examine how students in secondary education solve information problems and especially to determine the criteria students use while evaluating search results, sources and information using Internet for finding information while accomplishing an assignment. A description of what an information problem is will be given first.

An information problem arises when a discrepancy occurs between information needed to answer a question and information already known. An example of a simple information problem is: You have to travel from London to Harwich on September 8. You need to arrive at 18.00. At what time does your train leave? This problem can be solved easily by visiting a site of the UK railways. There is probably only one correct answer to this question. The assignment: "Write an essay on the treatment of cancer" is more complex. Solving this problem requires more than visiting only one webpage. There are several forms of cancer for instance and several treatment methods. Hence, more information has to be dealt with in this complex information problem and writing an essay is more difficult than answering one question.

Solving information problems requires several activities and these activities together constitute a process that we refer to as Information Problem Solving (IPS). The process

consists of the constituent skills *defining the information problem* (i.e. reading the task, activating prior knowledge), *searching information* (i.e. choose search strategy, specify search terms, evaluate search results), *scanning information* (i.e. read information global, evaluate source and information, elaborate on content), *processing information* (i.e. read in depth, evaluate information, store relevant information, elaborate on content) and *organising and presenting the information* (i.e. structure relevant information, realize the product).

Within the IPS process several problem solving processes can be distinguished. In the second phase, search information, one has to choose a search strategy and specify search terms. Choosing a search strategy and specifying search terms can be seen as problem solving processes on their own. What is the best strategy and best search term for the information problem at hand? With regard to the search strategy one could decide to use a depth first or breadth first strategy. Choosing a search term could be solved with the problem solving strategies brainstroming or trial and error, for instance. Another problem solving process within the IPS process, and the process that will be the focus of this article is evaluating sources and information. In the next section the importance of evaluating will be discussed.

Information problems can be solved with information from various sources. In the last decade, the World Wide Web (WWW) has become one of the largest and best accessible sources of information. And although there are many (more reliable) sources to collect information (e.g. the library or an encyclopaedia) the most probable source a student would use is the WWW (Beljaarts, 2006). Searching the Web for information differs from searching a library database or a table of contents. The web does not have an index, the amount of information is enormous and there are no gatekeepers that filter information. Research has shown that young children, university students and adults encounter difficulties when searching for information on the Web (Brand-Gruwel, Wopereis, & Vermetten, 2005; Duijkers, Gulikers-Dinjens, & Boshuizen, 2001; Hirsch, 1999; Kafai & Bates, 1997;

MaKinster, Beghetto & Plucker, 2002; Monereo, Fuentes & Sànchez, 2000). A problem for every age group is evaluating search results, and evaluating and selecting sources and information (Duijkers et al., 2001; Fidel, Davies, Douglass, Holder, Hopkins, et al., 1999; Hirsch, 1999; Kafai & Bates, 1997; Koot & Hoveijn, 2005; Lorenzen, 2002; Lyons, Hoffman, Krajcik, & Soloway, 1997; MaKinster et al., 2002; Wallace, Kupperman, Krajcik, & Soloway, 2000).

Evaluating results (which site am I going to open) is mostly done based on titles and summaries provided by the search engine, the number of results and the order of results (Duijkers et al., 2001; Fidel et al., 1999; Hirsch, 1999; Kafai & Bates, 1997; Koot & Hoveijn, 2005; Lyons et al., 1997; Wallace et al., 2000). Students' prior knowledge influences the criteria used to evaluate results. Students with a lot of prior knowledge evaluate results by title, origin, summary, and identifiers in the URL (.edu or .gov) (Makinster, 2002).

Evaluating and selecting sources and information (which information form which site am I going to use?) is not always done based on clear criteria (e.g., I see this is the official website of the National Health Organisation), but on intuition (e.g., I guess it's ok) (Koot & Hoveijn, 2005). Young children tend to believe that everything on the Web is true (Hirsch, 1999; Schacter, Chung & Dorr, 1998), especially when they find the same information on more sites (Koot & Hoveijn, 2005). Teenagers use information that can solve their information problem without thinking about the purpose of a site (Fidel et al., 1999). They also find it hard to express how they evaluate and select information (Lorenzen, 2002), and 82 % of the students admit that they rarely check information (Beljaarts, 2006).

An explanation for this lack of evaluating could be that students do not have enough metacognitive skills. As said before, problem solving skills call upon metacognitive skills (Perkins & Salomon, 1989; Van Merriënboer, 1997;1999). Research has indeed shown that students rarely use metacognitive strategies in the correct way (Bannert, 2004; Gerjets &

Scheiter, 2003; Rouet & Eme, 2002; Stadtler & Bromme, 2004).

So, it can be concluded that students hardly evaluate results, information and sources.

However, little research has been done on criteria students mention while solving information problems and if these criteria depend on the domain of the problem or if they are general. This study tries to fill this gap.

The research question is as follows:

How do students solve information problems, and what kind of criteria do students use when evaluating sources and information while searching the WWW for information?

This general question can be divided into several sub questions:

- o How much time do students spend on the constituent skills search, scan, process and organize while searching the World Wide Web and how often to they use these skills?

- o How often do students evaluate their search results, sources, and the information on a website?

- o Which criteria do students use for evaluating search results, sources, and information on a website?

- o Are the criteria general or do they depend on a certain domain?

- o Does students' evaluation behaviour influence task performance?

- o Are students aware of the criteria they use or can use?

Method

Participants

Twenty-three students of two schools for secondary education participated in this study. All students (8 boys and 15 girls; mean age 14.22, SD 0.422) were in the 9th grade of secondary education (Dutch: VWO, 6 year program).

Material

    Tasks. Teachers of the two participating schools created twelve information problems (4 physics, 4 geographic and 4 language and culture tasks). The problems consisted of a question, which had to be answered with information found on the Web. To answer this question, students had to choose which information to use and evaluate if the information could help solve the problem. To ensure comparability between problems, teachers received instruction on how to design them. For instance, it should not be a fact-finding question, the keywords leading to usable sites should not be mentioned, and solving the problem should require a search on the Web for preferably 30 minutes or more.

    Furthermore, to ensure comparability even more, constructs that define the difficulty levels of prose-task processing- type of information requested, type of match, and plausibility of distracters (Mosenthal, 1998) - were controlled for. *Type of information requested* refers to the kind of information that readers must identify. Examples are persons, amounts, goals, cause, effect (or outcome, result), evidence, opinion, explanation, equivalence and difference. Tasks that require the identification of cause, effect, equivalence and difference are more difficult to answer. *Type of match* refers to "the processes used to relate information in a question or directive to corresponding information in a text and to the processes used to select an answer from a range of response options" (Mosenthal, 1998, p.285). Type of match has several levels. The first level (locate tasks) requires students to match one or more features in a question to one or more features in a text. An example of a locate task is "how many Democrats have been president of the United States of America between 1980 and 2007? " This question gives the student information about the answer that is needed, the number of democratic presidents between 1980 and 2007. With tasks of the fourth level (generate tasks) students are required to make inferences, match given information in a question to corresponding information in a text and identify the correct requested information. Not all

information is in the question. An example of a fourth level task is "Discuss whether the title of the article represents a good summary of the article's main point". Students do not know how many points they need to discuss and they have to decided for themselves what the main point of the article is. The tasks used in this study are fourth level tasks: Usable keywords are not mentioned in the tasks and the tasks do not specify the number of responses required. *Plausibility of distracters* refers to possible different interpretations or definitions of features from a question or information. An example: when looking for information on the painter Francis Bacon, you come across a very informative website with information about his life and death. After reading it, you discover that this is a site about Francis Bacon, the philosopher. With a high plausibility of distracters a task becomes more difficult to solve. The tasks in this study were tested by using several keywords. With each keyword a number of sites came up that had the keyword either in the title or the summary but did not include usable information.

After the tasks were designed, a panel of four persons tested them. They solved the problems and filled out a questionnaire after each task, concerning time on task, keywords and websites, and difficulty level. Also the construct 'type of information requested' defined by Mosenthal (1998) was tested by the panel. Panel member 1 solved tasks 1-7, member 2 solved 1-12, member 3 solved 1-6 and member 4 solved 7-12. Table 1 summarizes the results for difficulty level (5-point likert scale: simple to difficult), type of information and time on task and Table 2 shows the mean difficulty level and SD for all tasks and per domain, and the mean and SD of time on task of all tasks and per domain.

Insert Table1 and 2 about here.

Table 1 shows some small differences between panel members on time on task and difficulty level, but all panel members agree on type of information requested. Difference in time on task can be explained by differences in prior knowledge of panel members. A

univariate analysis of variance on difficulty level and time on task showed no significant differences between the domains. Based on the remarks of the panel, some tasks were adjusted; the task on Britain's colonial history was removed, because it was found unclear and substituted by a task on MSN and SMS.

Table 3 presents the 12 tasks as used in the study. All tasks were preceded by the following text: "In this task you have to answer a question. You can search for information you need to answer the question on the Internet, and you can copy and paste the information in a word-file. Finally, try to answer the question in a few sentences, using the information you found. You have 30 minutes for this task. Think aloud during your search."

Insert Table 3 about here.

WWW knowledge Questionnaire. A questionnaire was used to obtain student's knowledge and conceptions about the WWW. The questionnaire consisted of three parts. The first part obtained additional information on name, age, sex, nationality, grades on Dutch, English, French, physics and geography. In the second part students were asked how much time they spent on the WWW per day, what they used the WWW for, and what their three favourite websites are. The third part consisted of 13 7-point Likert-scale items. Six items asked about the importance and usability of the WWW (e.g., I think it is useful to learn how to work with the Internet), $\alpha = .84$. Seven items asked about the expectations about their own knowledge on the WWW (e.g., I am sure I can learn how to work with the Internet), $\alpha = .87$.

Post hoc group interviews. A day after students completed the tasks, they returned in groups of three. In total there were eight focus groups. These interviews were used to obtain student's knowledge and conceptions on criteria for evaluating sources and information.

Procedure

During the individual sessions participants first filled out the WWW knowledge questionnaire, and then received an instruction on how to think aloud. After this instruction

each participant conducted two tasks. With 23 students and 12 tasks, this means that -accept

for the dialect task- all tasks were made four times. The dialect task was solved only twice.

After reading the first task, participants had to write down what they already knew about the

topic. Then, while thinking aloud, they searched the WWW for a maximum of 30 minutes to

solve the problem. Information found could be stored in a Word-file. The second task

followed the same procedure. The images of the computer screen and the audio were recorded

on videotape. Extra audio files were created with a laptop and the program Audacity. After

finishing the second task, participants were asked to verbalize their search procedure. (E.g.

"Could you tell me how you searched to solve this last task?" "Which things did you pay

attention to while searching?"). Participants returned in focus groups the next day for

approximately one hour. In every group three tasks were discussed (one geography, one

physics, one language and culture). This means that two students per group made each task.

They received a booklet with screen shots of websites they had opened and had or had not

used for their answer. They were asked to write down why they had or had not used the sites.

This was followed by a group discussion. First, students were asked what they had written

down and what they paid attention to while searching for information. This was done per task.

Then a general discussion followed, where students were asked to come up with criteria they

found important while searching. If students could not think of more criteria they were given

a few criteria and asked whether they thought they could be important. They were also asked

if they had paid attention to these criteria while searching the day before or if they ever paid

attention to these criteria. Finally, students were asked if the criteria were the same for the

courses language and culture, geography and physics. These sessions were also recorded.

Data analysis

All *think-aloud protocols* were transcribed verbatim. Then, all utterances and actions

were coded. The coding scheme to analyse the thinking aloud protocols was developed in

earlier studies (Brand-Gruwel et al., 2005; Wopereis & Brand-Gruwel, 2005). An inductive-deductive method was used to develop this system. This means that the coding system has an empirical and theoretical grounding; it has been tested and re-adjusted in a few iterations. For scoring the protocols two kinds of codes were used: descriptive codes and interpretative codes (Miles & Huberman, 1994). Descriptive codes entail little interpretation and can be linked to segments of the text in a straightforward way. Interpretative codes require more interpretation by the rater. The scoring system itself consisted of three types of categories. The first category, the constituent skills, was scored in an exclusive and exhaustive way. Also the time invested in the constituent skills was recorded. The second category consisted of the sub skills of each constituent skill. The third category entailed regulation activities. Overall, the variables measured with this coding system are the constituent and sub skills, and regulation activities of the Information-Problem Solving using Internet Model (IPS-I model) based on Brand-Gruwel et al., (2005). The sub skills evaluate search results, evaluate information and evaluate source were enriched for this study with criteria derived from literature on evaluating websites and information (Barker, 2005; Boekhorst, 2000; Beck, 1997; Ormondroyd, 2004; Kirk, 1996). Table 4 shows the constituent and sub skills and Table 5 shows the evaluation criteria of the coding system. Table 4 mentions all sub skills of the entire IPS process, in this article only the sub skills evaluate search results, evaluate information and evaluate sources are discussed.

Insert Table 4 and 5 about here

Two raters individually scored 6 of the 46 protocols. Interrater reliability was calculated. The similarity between the two raters, expressed in Cohen's Kappa was for all the protocols higher than .80.

The *task performance* (answer on the question) of the participants was judged based on three criteria: 1) Answer contains information of more than one source. Zero points for only

one site, one point for more than one site, 2) Quality of the sources is judged based on the judging criteria in Table 4. Quality can be zero to four points, 3) Suitability of information used was determined by identifying content related elements. Suitability can be zero to four points. Maximum score for task performance is nine.

The *post hoc group interviews* were also transcribed verbatim. Criteria mentioned by the students were grouped and labelled, based on the criteria used in the coding system (see Table 5). The criteria were grouped in four categories. The first category consisted of criteria mentioned by students, and the second of students' reactions to criteria mentioned by the researchers. The third category covered students' reactions to the question if the use of criteria differed between domains and the last category consisted of students' reactions to the question about the difference in use of criteria during the tasks and the use of criteria while searching in general.

To answer the research questions one-way ANOVAs were performed to determine if there was a difference in the amount of time spent on constituent skills, the frequencies of the used constituent skills, the frequencies of the sub skills evaluate results, evaluate information and evaluate source between the domains (language and culture, geography and physics). One-way ANOVAs were performed on the evaluation criteria, the total product score, the use of more than one site, the quality of sources, and suitability of information used to test if there was a difference between domains. To control for task order, one way ANOVAs on the amount of time spent on constituent skills, the frequencies of the used constituent skills, the frequencies of the sub skills evaluate results, evaluate information and evaluate source were also performed to determine the difference between the first and second task. These were all not significant. All ANOVAs were performed with an α of .05. The ANOVAs are performed to test separate hypotheses, therefore a Bonferroni correction is not necessary (Tutzauer, 2003).

Results

First analyses were performed to determine whether students were comparable in terms of

Internet experience. The majority of the participants (82,6%) spent 1-5 hours a day on the

Internet. The other 17,4% spent less than 1 hour a day on the Internet. The mean score on the

importance and usability of the WWW was 6.3, SD 0.70 and the mean score on the

expectations about their own knowledge on the WWW was 5.5, SD 0.80. This means

students find the WWW very usable and important to work with and expect that they know

enough about the WWW to be able to work with it.

Students' information problem solving processes

Figure 1 presents the average amount of time (in percentages of total time spent)

students spent on searching, scanning, processing and organizing while performing the tasks.

The constituent skill "Define information problem" is not included since all students were

asked to read the task and write down prior knowledge before starting their search. This

prescribed action that can be seen as defining the problem is not scored.

Insert Figure 1 about here

Figure 1 shows that students spent 44 % of the time on searching, 31 % on scanning,

16% on processing and 9 % on organizing information. Table 6 shows the means and standard

deviations per constituent skill for all tasks and for tasks per domain.

Insert Table 6

Two-way ANOVAs with domain and task order as between factors and time as a within factor

with four levels, were performed to test if the amount of time spent on each constituent skill

differed within subjects and between geographic, physics and language and culture tasks and

between the first en second task. Results showed no significant interaction between task order

and domain. There was a significant difference within subjects on the amount of time spent on

each constituent skill, $F(3,120)=23.51$, MSE=10796.50, $p=.00$, $\eta^2=.37$. Students spent significantly more time on searching than on scanning, $F(1,40)=5.30$, MSE=6438.03, $p=.03$, $\eta^2=.12$. They spent significantly more time on searching than on processing $F(1,40)=26.05$, MSE=34596.37, $p=.00$, $\eta^2=.39$, and organizing $F(1,40)=57.30$, MSE=52939.53, $p=.00$, $\eta^2=.59$. Students also spent significantly more time on scanning than on processing $F(1,40)=10.77$, MSE=11185.95, $p=.00$, $\eta^2=.21$, and organizing $F(1,40)=46.56$, MSE=22454.59, $p=.00$, $\eta^2=.54$. There was no significant difference between the amounts of time spent on the skills process and organize.

The IPS process turned out to be iterative and students switched frequently between the constituent skills. After an initial search, they either selected a site from the hit list, or they adjusted their search term and started another search. This can be seen in Table 7. If participants would follow the cycle search-scan-process-organize-search the frequencies would be equal. Since the frequencies of search and scan are high, it can be assumed that many iterations are made between these processes. Table 7 shows the means and standard deviations of the used constituent skills (frequencies) calculated over all 46 tasks (23 students * 2 tasks) and on the 4 tasks per domain.

Insert Table 7 about here

Two-way ANOVAs with domain and task order as between factors and frequencies as a within factor with four levels, were performed to test if the use of constituent skills differed within subjects and between domain and task order. Results showed no significant interaction between task order and domain. There was a significant difference within subjects on the use of each constituent skill, $F(3,120)=97,39$, MSE=1921.38, $p=.00$, $\eta^2=.71$. Students searched significantly more than scanned $F(1,40)=19.60$, MSE=132.07, $p=.00$, $\eta^2=.33$, processed $F(1,40)=116,32$, MSE=5954.78, $p=.00$, $\eta^2=.75$, or organized $F(1,40)=135.69$, MSE=7208.01, $p=.00$, $\eta^2=.77$. Students also scanned significantly more than processed $F(1,40)=69.04$,

MSE=4313.21, *p=.00, η²= .63* or organized F(1,40)=99.81, MSE=5388.70, *p=.00, η²= .71.*

Finally, students significantly processed more than they organized F(1,40)=6.32, MSE=59.80,

p=.02, η²= .14.

So, it can be concluded that the process is iterative, especially between the search and scan processes. The fact that so many iterations are made between these processes makes it even more interesting to determine the criteria students use while evaluating. How do students decide to select a certain site or to click it away? This process of evaluating occurs quite often, because of the reasonable number of iterations. In the next section results on students' evaluating processes are presented.

Students' evaluating processes

To answer the research questions how often students evaluate during the IPS process, what criteria they use, and whether these criteria are domain specific, the protocols were analysed in depth. Table 8 shows the means and standard deviations of the frequencies of the sub skills 'evaluate search results' (i.e., evaluate the hit list), 'evaluate information' and 'evaluate source'.

Insert Table 8 about here

It should be noted that only the evaluations that were explicitly mentioned were scored. Students also clicked pages away without telling why. It those cases, it would seem they did evaluate the site, but it is not clear on what grounds. These evaluations are not included in Table 8.

It seems that students hardly evaluate sources and information explicitly (neither during scanning the site nor during processing). Especially, sources are hardly evaluated. Evaluating the hit list to select a site that may give an answer to the problem is done most frequently. Furthermore, standard deviations are large, especially on the skills 'evaluate search results'

and 'evaluate information while scanning sites'. Differences between students are large concerning the execution of these skills.

Two-way ANOVAs with domain and task order as between factors and kind of evaluations as a within factor with five levels were performed to test if the frequencies and kind of evaluations (evaluate search results, evaluate information during scan, evaluate information during process, evaluate source during scan, evaluate source during process) differed within subjects and between domain and task order. Results showed no significant interaction between task order and domain. There was a significant difference within subjects on the use of evaluations, $F(4,160)=69.54$, MSE=726.83, $p=.00$, $\eta^2=.64$. Participants significantly more often evaluated search results than information during scanning, $F(1,40)=34.92$, MSE=1627.21, $p=.00$, $\eta^2=.47$ and than source during scanning, $F(1,40)=89.51$, MSE=4000.79, $p=.00$, $\eta^2=.69$. Search results were also significantly evaluated more than the information during processing, $F(1,40)=77.58$, MSE=3293.60, $p=.00$, $\eta^2=.66$, and the source during processing, $F(1,40)=91.89$, MSE=4139.47, $p=.00$, $\eta^2=.70$. Participants significantly more often evaluated information during scanning than the source during scanning, $F(1,40)=65.93$, MSE=525.02, $p=.00$, $\eta^2=.62$, and information during processing, $F(1,40)=35.74$, MSE=290.75, $p=.00$, $\eta^2=.47$ and the source during processing $F(1,40)=65.46$, MSE=576.00, $p=.00$, $\eta^2=.62$.

Participants significantly more often evaluated source during scanning than the information during processing, $F(1,40)=13.32$, MSE=34.36, $p=.00$, $\eta^2=.25$, and source during processing, $F(1,40)=6.74$, MSE=1.18, $p=.01$, $\eta^2=.14$. Finally, the information during processing was significantly evaluated more than the source during processing $F(1,40)=18.57$, MSE=48.29, $p=.00$, $\eta^2=.32$.

Evaluation criteria

Although students hardly gave expression to their evaluation, it is important to determine which criteria they used when they evaluated results, information and sources. Table 9 shows the criteria as defined in the method section and the frequencies on how often students used these criteria for evaluating results, sources and information.

Insert Table 9 about here

The students appeared to use only a few of the possible 29 criteria during information problem solving. They mainly evaluated search results or selected sites in the hit list based on title and summary. They read the title and the summary and decided if the information on the site could help them find an answer on the question and then decided to open the site or not. Only a few times was the rank in the hit list or language used as criteria. In total (46 tasks) the results were viewed 949 times and they were explicitly evaluated 217 times, a percentage of 24 %. The following description gives an impression of students' evaluation behaviour.

A student has just performed a search with Google, and is looking at the results. He points his mouse to the summary, reads it aloud: "French influence on Dutch dialects" and opens this site.

The usability of the information found on a site was often evaluated with the criterion '*connection to task*'. This means that students sought for information that could help answer the question, so the criterion was the content. Sometimes the language (foreign, slang, spelling errors) was a reason for determining if information was usable. Criteria of verifiability were hardly used to evaluate information. On two occurrences a student connected the information to prior knowledge. Reliability of information was checked eight times with the criterion '*kind of information*'. Information was evaluated 117 times out of

780 page views (15 %). The following description gives an impression of students' evaluation behaviour.

A student is trying to answer a task on French dialects. He opens a page, scans it, and while closing the page, says: "It doesn't say anything here about the dialects".

The source was evaluated on speed and one time on the kind of source. Only 4 times out of 780 page views the source was judged, a percentage of only 0.5%. The following utterances give an impression of students' evaluation behaviour.

The following description gives an impression of students' evaluation behaviour.

A student has opened a page and it is loading. After a minute, only the coloured background has appeared. The student says: "This takes too much time" and closes the page.

Students also evaluated results, information and sources without mentioning criteria. Utterances like "these are good results", "this is a nice website", are examples of these kinds of evaluations. In total there were 316 of these undefined evaluations. Since the focus in this article is on criteria mentioned by students, these undefined evaluations are not included in the rest of the result section. We will come back to them in the discussion.

Criteria for evaluating: domain specific

Table 10 presents the means en standard deviations of the used criteria per domain.

Insert Table 10.

Two-way ANOVAs with domain and task order as between factors and used criteria search results as a within factor with four levels, two-way ANOVAs with domain and task order as

between factors and used criteria information as a within factor with six levels and two-way ANOVAs with domain and task order as between factors and used criteria source as a within factor with two levels were performed, to test if the use of criteria differed within subjects and between domain and task order. Results showed no significant interaction between task order and domain. There was a significant difference within subjects on the use of search results criteria, $F(3,120)=33.67$, MSE=242.88, $p=.00$, $\eta^2= .46$.

The search results were significantly judged more on "*title/summary*" than on "*rank in hit list*", $F(1,40)=33.04$, MSE=967.78, $p=.00$, $\eta^2= .45$, "*known to user*", $F(1,40)=34.39$, MSE=966.68, $p=.00$, $\eta^2= .46$, and "*language*" $F(1,40)=33.99$, MSE=980.02, $p=.00$, $\eta^2= .46$.

There was a significant difference within subjects on the use of information criteria, $F(5,200)=34.82$, MSE=25,87, $p=.00$, $\eta^2= .47$.

The information was evaluated significantly more on "*connection to task*" than on "*language*" $F(1,40)37.33$, MSE=137.03, $p=.00$, $\eta^2= .48$, "*amount*" $F(1,40)=41.10$, MSE=148.53, $p=.00$, $\eta^2= .51$, "*information agrees with prior knowledge*", $F(1,40)=43.55$, MSE=167.79, $p=.00$, $\eta^2= .52$, "*organization*", $F(1,40)=43.45$, MSE=171.03, $p=.00$, $\eta^2= .52$ and "*kind of information*", $F(1,40)=45.00$, MSE=144.64, $p=.00$, $\eta^2= .53$. Information was also evaluated significantly more on "*amount*" than on "*information agrees with prior knowledge*", $F(1,40)=6.79$, MSE=0.59, $p=.01$, $\eta^2= .16$, and "*organization*", $F(1,40)=5.00$, MSE=0.79, $p=.03$, $\eta^2= .11$. Information was evaluated significantly more on "*language*" than on "*organization*", $F(1,40)=5.25$, MSE=1.88, $p=.03$, $\eta^2= .12$.

There was no significant difference in the use of criteria to evaluate sources.

Students' products

Evaluating is crucial when selecting information, since the World Wide Web contains so much information that is incorrect, subjective or biased. If one uses information for an essay

or answer to an information problem without questioning the source or considering if the information is correct it is possible that the end product is not optimal. In this section, the quality of the students' answers to the questions asked in the tasks are analysed and related to used criteria.

To answer the questions about the quality of the solved problems and the relation to the used criteria, students' answers were analysed. Table 11 shows the means and standard deviations of the total product score and the scores on quality of sources and suitability of the information.

Insert Table 11 about here

The mean scores on all the aspects are low. The maximum possible total product score is 9 and the mean is 4.1. The maximum achieved total score is 7. The maximum (possible and achieved) score on quality of the answer is 4 and the mean is 1.6. Thus, student scores are far from optimal.

A one-way ANOVA shows that the subject area (geography, physics and language and culture) has a significant influence on the total product score, $F(2,45)=5.23$, MSE=12.22, $p=.01$, $\eta= 0.20$ and the suitability of information, $F(2,45)=10.23$, MSE=9.26, $p=.00$, $\eta= 0.32$. The four physics tasks resulted in significantly higher scores than the task from other domains. Posthoc analysis showed that there was no significant difference between the four physics tasks. One-way ANOVAs on differences between the quality of the sources and the score on the use of more than one source did not yield significant results.

Regression analyses with the number of evaluations of search results, information and sources as predictors and the total product score and suitability of information as dependent variables were not significant. Regression analyses with the used criteria as predictors and total product score and suitability of information as dependent variables were not significant either.

Students' awareness of evaluation criteria

As shown, students do not use a lot of criteria for evaluating results, information and sources. Is that because they are not aware of these criteria or because they do not feel the need to use or explicitly express them? In this paragraph the result of the group interviews will be discussed.

Insert Table 12 about here

Table 12 shows the criteria and the number of groups that mentioned these criteria during the group interviews. Only two criteria are mentioned by all eight groups; *connection to task* and *information agrees with more sites*. More than half of the groups mentioned criteria like *title/summary, language* and *appearance.* Four or fewer groups mentioned criteria like *audience*, *author* and *goal*. Comparing Table 12 to Table 9 (the criteria actually used during the tasks) makes it clear that students mentioned more criteria during the group interviews than they explicitly used during the tasks. None of the students mentioned *information agrees with more sites* during the tasks but this criterion is mentioned during all interviews. Criteria like *audience*, *author*, *references* and *goal* are not mentioned while solving the tasks, but are mentioned by some groups. Criteria like *topicality*, *objectivity* and *primary/secondary* are neither used nor mentioned.

The results on students' use of criteria are nuancated by what the students seem to know about criteria when asked about criteria. They seem to know more criteria than they explicitly use. This difference could be explained by the fact that students do not only use criteria they know (and explicitly mention when evaluating) but are sometimes triggered to use a criterion. When triggered to use a criterion students are not always aware of using the criterion, and do not mention it explicitly, but when asked about the criterion later, they do have knowledge about this criterion. For instance the criterion *author* is only important if the

author is mentioned on the website. When the author is not mentioned, they do not try to find out who the author is and it is not a reason to doubt the information. This could imply that the criterion author is not a criterion that is used systematically by students, but its use is triggered. However, this triggered use, could lead to use of the criterion based on misconceptions. One student said he did not check for the author because the author was not mentioned and he did not know the website. This implies, that he would only check the author if the author was mentioned. In fact, correct use of the criterion would be to doubt a website that does not mention an author. To some students, the author is not important, as long as they understand the information and the information agrees with more sites.

Their view on the use of the criterion *information agrees with more sites* is rigid. If two sites contain the same information, this information is regarded as true. Of course, this way of using the criterion could lead to the use of false or biased information, since some websites copy content of other sites. If the source is not checked by criteria like author, references and goal, the information could agree with more sites, but may still be false or biased. Students do not always realize this, as shown by the following remark: "They made the effort to build a page, why would they put information on it that is not true?"

Furthermore, students were asked if the criteria they did not mention could be important. The criterion *topicality* was an eye-opener for most students. They admitted they hardly looked at the date a page was last updated. They agreed that *topicality* could be important, depending on the information problem. "If you have to find out where New York is located, that will be the same for a couple of years." The suitability of the other criteria also depended on the task and on the students' perceptions of teachers' expectations. As for the criterion *kind of information*, students admitted that they knew that information from a forum or an opinion of somebody is not always reliable, but if they cannot find anything else they will use the information anyway.

All students felt that the basic criteria (*connection to task* and *information agrees with more sites)* were important no matter what the domain was. The other criteria depended on the task, and not really on the domain, in their view.

The overall image created by these group interviews was that the students opened websites based on the summary given by the search engine and checked the information on connection to task. The information had to be easily traceable and had to agree with more sites. The author was not really important; the organization behind a site could be a clue for the reliability. But if the organization was not mentioned, that was no reason to discard the information, especially when the information was found on more sites. Dutch sites were preferred and the information should be understandable.

Discussion

The central question of this article is how students solve information problems and what kind of criteria they use when evaluating results, sources and information while searching the WWW for information.

While solving information problems students spent most of their time on searching and scanning and only a small amount of time on processing and organizing information. This finding is in line with results Brand-Gruwel et al. (2005) found. Ph-D-students and freshman also use more time on searching and scanning than on processing. Subjects in their study however, spent a substantial amount of time organizing and presenting the information, but this was due to the fact that they had to deliver an outline for an article.

During the IPS process, students do not very often evaluate results, information and sources. Only 24% of the times that search results are viewed they are evaluated explicitly, information is evaluated in only 15 % of the page views and sources are evaluated only 0.5 %

of the views. When students evaluate the results of a hit list, they mostly use the title and the given summary. When evaluating information the criterion used most often is if the information is connected to the task. The criteria students mention, when asked which are important for evaluating information, are not always the same criteria they mention while solving the information problems. During the group discussion the criterion if the information can be found on more sites was mentioned as important. They mentioned more criteria but also admitted that they did not always use them while searching the Web. The reasons were time pressure, motivation and convenience.

Furthermore, the students did not mention differences between the domains. All criteria are applicable in the different domains, although type of task can make criteria less or more important.

If students do not evaluate search results, information or sources, one may expect a less optimal product. Results show that students' task performance was far from optimal, and that the quality was rather low. However no relation between evaluation behaviour and task performance could be found. Students who use more criteria do not achieve significantly higher results. However, this could be due to the fact that although they used more criteria, they did not always use the more sophisticated criteria or used the criteria in a correct way.

One of the limitations of this study was the time pressure students felt during the experiment. Some students admitted they did not check information because of the time limit of 30 minutes. They felt they had to come up with an answer and were glad when they found information. They said they would pay more attention to the criteria when they had to write an essay or when they were searching for something they were really interested in. This could mean they would achieve better results in those cases.

However, there is certainly room for improvement, because there are still criteria they had never thought about and it is important to make them aware of their use of criteria. It is not only important for schoolwork, but in every day life as well. Most striking was the fact that students expected to find an answer on one single website, preferably in the first couple of sentences. They did not bother to combine information or check on things. If the author of the site was not mentioned they did not try to find out who the author was. Students wanted the information served on a silver platter and did not want to do a lot of work themselves.

Another factor that may have influenced the results is the data collection procedure. Thinking aloud is not always easy for students. They discarded sites without explaining why and as mentioned before there were 316 utterances with undefined criteria. This could be an explanation for the difference between the used criteria and the criteria mentioned in de focus groups. Research by van Gog, Paas, Van Merrienboer and Witte (2005) showed that cued retrospective reporting (the original task and a record of eye movements is used as a cue for retrospection) worked better than concurrent and retrospective reporting while trouble shooting with electrical circuits. Participants provided more action information, 'how' information and metacognitive information. Perhaps this method could also work for investigating information problem solving processes and the use of criteria. The search could be discussed in retrospect with the students and students could see the eye movement data to remember what they were looking at. Perhaps this would result in a better explanation of why they discarded some sites in only a few seconds.

The results of this study agree with earlier studies in this field. Students have trouble evaluating information and do not have a critical attitude towards information on the WWW. Koot and Hoveijn (2005) also used think aloud sessions and interviews in their study of 11-year olds. Their focus was also on the children's view on the reliability of information. These young children approached the Web the same way as the high school students in our

study. They also hardly evaluated results, sources and information. They were aware that not all information on the Web is reliable, but seemed to forget this while searching. And just like our students, there was a discrepancy between what they did and what they reported. When asked, they were capable of mentioning rational arguments why information was reliable, but when searching they based their evaluations more on intuition. Lorenzen (2002) interviewed high school students about how they used the World Wide Web to find information for school assignments, how they went about finding information and how they knew if the information they found on the Web was scholarly or factual. Students admitted that they were not really sure how they could distinguish between good and bad information. They tended to trust institutional pages, knew to check spelling, bibliography and web pages author. However, they seemed to forget that the authors name on a site does not mean this person is real or has actually written the information on a site. It is also possible that a site may look like an institutional page, but it is not. A big problem was that students evaluated a site based on how elaborate it looks: "if a web site looks good, appears to be professional, and has a lot of detail on it, many of the students will accept it as a good web site for information. Clearly, this method of evaluation is weak and not really helpful" (Lorenzen, 2002, p. 161). The students in the study by Lorenzen (2002) seemed to approach the web in the same way our students did.

Finally, a poll by a research company (Beljaarts, 2006) also showed that students felt they could trust the Internet and hardly checked their information. The students in our study were not different from students in other age groups or students of their own age.

This attitude (not evaluating information critically) can cause problems in school and everyday life. It is important to make students aware of their evaluation behaviour, the misconceptions they may have and point them to the criteria they can use to evaluate

information. This will help them achieve better results in school. Future research should aim on developing instruction in IPS, focusing on evaluating and selecting. Since IPS is a complex cognitive skill, the students should work on whole tasks, which are authentic and comprehensive. These tasks require from students to perform all the constituent skills that make up the whole complex skill during task performance (van Merrienboer, 1997). And since IPS is important throughout life, instruction should also focus on the transfer of the skill to multiple domains, tasks and situations.

References

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.

Barker, J. (2005). *Critical evaluation of resources.* Retrieved May, 16, 2006, from UC Berkeley Library Web site http://www.lib.berkeley.edu/TeachingLib/Guides/Evaluation.html

Beck, S. *"Evaluation Criteria." The Good, The Bad & The Ugly: or, Why It's a Good Idea to Evaluate Web Sources.* (1997). Retrieved May 10, 2006, from http://lib.nmsu.edu/instruction/evalcrit.html

Beljaarts, M. (2006). *Internet, een populair medium voor het zoeken van informatie bij schoolopdrachten [Internet, a popular medium for searching information for school assignments].* 's-Hertogenbosch: Malmberg.

Boekhorst, A. K. (2000). *Informatievaardig worden in het onderwijs, een informatiewetenschappelijk perspectief: Een vergelijkende gevallenstudie in Nederland en Zuid-Afrika [Becoming information literate in education, an information science perspective: a comparative case study in The Netherlands and South Africa].* Unpublished dissertation. Retrieved May 10, 2006 from http://cf.hum.uva.nl/akb/publicaties/prom-akb-tot.pdf

Brand-Gruwel, S., Wopereis, I., & Vermetten, Y. (2005). Information problem solving: Analysis of a complex cognitive skill. *Computers in Human Behavior 21,* 487-508.

Bransford JD, Brown AL, Cocking RR (Eds.) (2000) *How People Learn: Brain Mind, Experience, and School.* National Academy of Sciences. National Academy Press, Washington, D.C.

Duijkers, H. M., Gulikers-Dinjens, M. T. H., & Boshuizen, H. P. A. (2001). Begeleiden van leerlingen bij het zoeken, selecteren en beoordelen van informatie. [Supporting students when searching, selecting and judging information] In: J. Ahlers, T. Hoogbergen, P. Leenheer, & J. de Vos (Eds.), *Handboek Studiehuis Tweede Fase*. Alphen a/d Rijn: Samsom.

Eisenberg, M. B., & Berkowitz, R. E. (1990). *Information problem-solving: The big six skills approach to library and information skills instruction.* Norwood, NJ: Ablex.

Fidel, R., Davies, R. K., Douglass, M. H., Holder, J. K., Hopkins, C. J., Kushner, E. J., Miyagishima, B. K., & Toney, C. D. (1999). A visit to the information mall: web searching behavior of high school students. *Journal of the American Society of Information Science, 50* (1), 24-37.

Gerjets, P., & Scheiter, K. (2003). Goal configurations and processing strategies as moderators between instructional design and cognitive load: Evidence from hypertext-based instruction. *Educational Psychologist*, 38, 33-41.

Goldstein F. C., & Levin H. S. (1987). Disorders of reasoning and problem-solving ability. In M. Meier, A. Benton, & L. Diller (Eds.), *Neuropsychological rehabilitation*. London: Taylor & Francis Group.

Grimes, D. J., & Boening, C. H. (2001). Worries with the Web: A look at student use of Web resources. *College and Research Libraries, 62*(1), 11-23.

Hirsch, S. G. (1999). Children's relevance criteria and information seeking on electronic resources. *Journal of the American Society for Information Science*, *50* (14), 1265-1283.

Kafai, Y., & Bates, M. J. (1997). Internet web-searching instruction in the elementary classroom: building a foundation for information literacy. *School Library Media Quarterly, 25* (2), 103-111.

Kirk, E. E. (1996). *Evaluating information found in the Internet.* Retrieved May, 10, 2006 from Johns Hopkins University, the Sheridan Libraries Web site http://www.library.jhu.edu/researchhelp/general/evaluating/index.html

Koot, U., & Hoveijn, M. (2005). Google, de magische antwoordmachine. [Google, the magical answermachine] *Pedagogiek in Praktijk, 11* (28), 18-22.

Lorenzen, M. (2002). The land of confusion? – High school students and their use of the World Wide Web for research. *Research strategies, 18* (2), 151-163.

Lyons, D., Hoffman, J., Krajcik, J., & Soloway, E. (1997). *An investigation of the use of the*

World Wide Web for on-line inquiry in a science classroom. Paper presented at the meeting of the National Association for Research in Science Teaching, Chicago, IL.

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* (2), 155-172.

Monereo, C., Fuentes, M., & Sànchez, S. (2000). Internet search and navigation strategies used by experts and beginners. *Interactive Educational Multimedia,* 1, 24-34.

Moore, P. (1995). Information problem solving: a wider view of library skills. *Contemporary Educational Psychology*, 20, 1-31.

Rouet, J.-F., & Eme, E. (2002). The role of metatextual knowledge in text comprehension. In P. Chambres & M. Izaute (Eds.), *Metacognition: Process, function and use* (pp. 121-133). Kluwer Academic Publishers.

Ormondroyd, J. (2004). *Critically analysing information sources.* Retrieved May 10, 2006, from Cornell University Library, Reference Department, Instruction, Research, and Information Services (IRIS), Ithaca, NY, USA Web site
http://www.library.cornell.edu/olinuris/ref/research/skill26.htm

Perkins, D. N., & Salomon, G. (1989). Are cognitive skills context-bound? *Educational Researcher,18,* 16-25.

Rothenberg, D. (1998). How the Web Destroys Student Research Papers. *Education Digest 63*, 59-61.

Schacter, J., Chung, G. K. W. K., & Dorr, A. (1998). Children's Internet searching on complex problems: performance and process analyses. *Journal of the American Society for Information Science, 49* (9), 840-849.

Simons, R.J., Van der Linden, J., & Duffy, T. (2000). New learning: three ways to learn in a new balance. In: R.J. Simons, J. Van der Linden & T. Duffy, (Eds.). *New learning* (pp 1 – 20). Dordrecht: Kluwer.

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.

Tutzauer, F. (2003). On the sensible application of familywise alpha adjustment. *Human Communication Research, 29*(3)*, 455-463.

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 Merriënboer, J. J. G. (1997). *Training complex cognitive skills*. Englewood Cliffs, NJ: Educational Technology.

Van Merriënboer, J.J.G (1999). *Cognition and Multimedia Design*. Inaugural address. Open

University of the Netherlands.

Wallace, R. M., Kupperman, J., Krajcik, J. & Soloway, E. (2000). Science on the web:

students online in a sixth-grade classroom. *The journal of the learning sciences, 9*(1), 75-104.

Wopereis, I, Brand-Gruwel, S. & Vermetten, Y. (2005, August). The effect of embedded
instruction on solving information problems. Paper presented at the 11th European Conference
for Research on Learning and Instruction (EARLI), Nicosia, Cyprus.

Table 1

Results of the Test Panel concerning Difficulty Level, Type of Information Requested and

Time on Task

Task (domain)	Panel member	Difficulty level	Type of information requested	Time on task (minutes)
1 Kyoto (geography)	1	Difficult	D	30
	2	Fairly difficult	D	15
	3	Fairly difficult	D	9
2 French dialects (language and culture)	1	Fairly simple	B/D	21
	2	Fairly simple	D	20
	3	Fairly simple	A/D	12
3 Great Britain colonial history (language and culture)	1	Difficult	D	30
	2	Difficult	D	30
	3	Fairly difficult	D	11
4 Road reflection (Physics)	1	Easy	D	30
	2	Easy	D	5
	3	Fairly Difficult	D	14
5 Coal (geography)	1	Difficult	D/E	30
	2	Difficult	D	25
	3	Easy	D	10
6 Resemblance English and French language (language and culture)	1	Fairly difficult	D	30
	2	Fairly difficult	D	20
	3	Easy	D	5
7 Hail (Physics)	1	Simple	D/B	30
	2	Easy	D	15
	4	Fairly difficult	D	10
8 Missing children (language and culture)	2	Difficult	D	30
	4	Difficult	D/C	20
9 Electron (Physics)	2	Easy	D	10
	4	Easy	D	11
10 Pollution and water (geography)	2	Fairly simple	D	15
	4	Fairly difficult	D	28
11 City patterns (geography)	2	Fairly difficult	D	30
	4	Difficult	D	17
12 Snow (Physics)	2	Fairly difficult	D	25
	4	Difficult	D	10

Note: Type of information requested:
A = Identification of person, group, location or object
B = Identification of amount, time, attribute, species, actions, locations or procedures
C = Identification of goal, function, alternative, attempt, condition, order, verification or problem
D = Identification of cause, effect, result, evidence, opinion or explanation
E = Identification of equivalence, difference, theme

Table 2

Means and Standard Deviations of difficulty level and time on task

Tasks	N	Difficulty level Mean (SD)	Time on task Mean (SD)
All	12	3.8 (0.9)	19.3 (4.0)
Geography	4	4.0 (0.7)	21.2 (2.3)
Physics	4	3.4 (0.8)	15.7 (3.5)
Language and culture	4	3.8 (1.3)	21.2 (3.7)

Table 3

The Tasks used in the Study

Number	Title	Course	Task (translated from Dutch)
1	Kyoto	Geography	Ecological changes have become clearly visible during the last years. The snowline of the Kilimanjaro is moving and the ice of the polar caps is melting. Can all this be prevented if the Kyoto protocol is followed?
2	Dialect	Language and Culture	There are several dialects within the French language. Which dialects exist and what are the differences between these dialects?
3	MSN & SMS	Language and Culture	Young people used msn and sms a lot. Does this affect their Dutch?
4	Road reflection	Physics	Under which circumstances does the road reflect like a mirror in the distance?
5	Coal	Geography	Why does the Netherlands have charcoal layers of 1-2 metres thick, situated far below the ground, while the USA has layers 40 metres thick situated near the surface of the earth?
6	French and English	Language and Culture	How can the equivalences between the English and French vocabulary be explained?
7	Hail	Physics	Why are hailstones sometimes small and sometimes big?
8	Missing children	Language and Culture	How effective is searching for missing children by placing their pictures on milk cartons?
9	Electron	Physics	An electron has a charge, but what is this charge exactly?
10	Pollution	Geography	What is the effect of pollution on the quality of tap water?
11	City patterns	Geography	You have probably never thought about it, but many cities have certain city patterns. European cities are built different that cities in the USA. What are the differences in city patterns?
12	Snow	Physics	Why is there so much air in a snowflake?

Table 4

Thinking Aloud Protocols: Constituent and Sub Skills Scored

Constituent skill	Sub skills
Define information problem	- read task
	- formulate questions
	- activate prior knowledge
	- clarify task requirements
	- determine needed info
Search information	- select search strategy
	- specify search terms
	- evaluate search results
Scan information	- read information global
	- evaluate source and scanned information
	- store relevant information
	- elaborate on content
Process information	- read in depth
	- evaluate processed information
	- store relevant information
	- elaborate on content
Organize and present information	- formulate problem
	- structure relevant information
	- outline the product
	- realize the product
	- elaborate on content

Table 5

Thinking Aloud Protocols: Scored Evaluation Criteria

Sub skill	Criteria	
Evaluate search results	1.Title/Summary	What is the title given by the search engine and what is the contents of the summary?
	2.Kind (site/PDF/)	What kind of source is it, a website, a word document, a PDF file?
	3.Address	What is the address? Is it a .com or .org address?
	4.Rank in hit list	How many results are there in total and what is the rank of the result I am evaluating?
	5.Known to user	Have I used this site before, or have I heard good or bad things about it?
	6.Language	Is the site in a language I prefer and/or understand?
Evaluate information	**A Usability**	
	1. Language	In what language is the information written? Are there many grammatical or type errors? Is it filled with domain specific language?
	2. Connection to task	Does the information answer (part of) the information problem?
	3. Audience	Is the information aimed at a specific group of readers?
	4. Topicality	Is the information up to date?
	5. Amount	Is there enough information on the page? Or only a part of the information I'm looking for?
	B Verifiability	
	1. Author	Who has written the information? Can I contact him/her?
	2. References	Are there references on the page to used sources? Or links to more websites on the same subject?
	3. Information agrees with more sites	Can I find the same information on more pages or is this information only available on this site?

	4. Information agrees with prior knowledge	Does the information confirm what I already know?
	5. Organization	Which organization is behind the information? A governmental organization or a health organization? Can I find their logo on this site?
	C. Reliability	
	1.Kind of information	What kind of information is it? A newspaper article or a forum? Is it an opinion or results from research?
	2. Objectivity	Is the information objective or coloured by a certain point of view? Are there a lot of advertisements on the page?
	3. Primary/Secondary	Is the information first hand or is it someone telling about someone who did something?
	4. Goal	What does the (author of) information want to achieve. Sell something? Convince me of something or just inform me?
Evaluate source	**A Technical**	
	1. Appearance	Does the site appeal to me? Does it have pictures or only text?
	2. Speed	Does it take a lot of time to load the page?
	B Usability	
	1. Language	In what language is the site written? Are there many grammatical or type errors?
	2. Connection to task	Does the site have a connection to (part of) my information problem?
	3. Audience	For who is the site meant? Who are its visitors?
	4. Topicality	Is the site updated regularly?
	C Verifiability	
	1. Reputation	Is this site famous or infamous for something? Does it have a good/bad reputation?
	D. Reliability	
	1. Kind (site/PDF)	What kind of source is it, a website, a word document, a PDF file?
	2. Primary/secondary	Is the site an original source or a site telling about what is

Table 6

Means and Standard Deviations of time spent on constituent skills

Tasks	N	Search Mean (SD)	Scan Mean (SD)	Process Mean (SD)	Organize Mean (SD)	Total Mean (SD)
All	46	10.7 (6.7)	7.0 (4.1)	3.9 (4.8)	1.7 (2.1)	23.9 (7.1)
Geography	16	11.3 (6.1)	8.2 (3.9)	4.2 (5.4)	2.2 (2.5)	26.5 (4.2)
Physics	16	9.4 (7.4)	4.8 (2.8)	4.3 (4.9)	1.4 (2.0)	20.4 (8.9)
Language and culture	14	11.6 (7.0)	8.4 (4.7)	3.1 (4.1)	1.4 (1.6)	25.1 (6.1)

Table 7

Frequencies of Constituent Skills

Tasks	N	Search Mean (SD)	Scan Mean (SD)	Process Mean (SD)	Organize Mean (SD)
All	46	14.5 (6.8)	12.8 (6.9)	2.9(2.9)	1.6 (1.8)
Geography	16	16.6 (7.5)	15.1 (8.7)	2.9 (3.0)	1.4 (1.4)
Physics	16	14.4 (7.4)	12.4 (6.3)	3.6 (3.4)	1.5 (2.1)
Language and culture	14	12.2 (4.8)	10.5 (4.2)	2.1 (2.1)	2.0 (1.9)

Table 8

Frequencies of Evaluating Search Results, Sources and Information per Task per Constituent

Skill

Tasks	N	Evaluate results during search Mean (SD)	Evaluate information during scan Mean (SD)	Evaluate source during scan Mean (SD)	Evaluate information during process Mean (SD)	Evaluate source during process Mean (SD)
All	46	9.5 (6.6)	3.5 (3.2)	0.2 (0.5)	1.1 (1.7)	0.0 (0.2)
Geography	16	10.3 (6.8)	4.3 (3.7)	0.2 (0.4)	1.3 (1.7)	0.0 (0.0)
Physics	16	9.6 (7.3)	3.3 (3.1)	0.2 (0.5)	1.4 (2.0)	0.1 (0.3)
Language and culture	14	8.6 (6.0)	2.9 (2.8)	0.2 (0.4)	0.5 (0.9)	0.0 (0.0)

Table 9

Frequencies of Criteria used for Judging Search Results, Sources and Information

Sub skill	Criteria	Frequencies
Evaluate search results	1.Title/Summary	217
	2.Kind (site/PDF/)	0
	3.Address	0
	4.Rank in hit list	3
	5.Known to user	4
	6.Language	2
Evaluate information	**A Usability**	
	1. Language	10
	2. Connection to task	89
	3. Audience	0
	4. Topicality	0
	5. Amount	7
	B Verifiability	
	1. Author	0
	2. References	0
	3. Information agrees with more sites	0
	4. Information agrees with prior knowledge	2
	5. Organization	1
	C. Reliability	
	1.Kind of information	8
	2. Objectivity	0
	3. Primary/Secondary	0
	4. Goal	0
Evaluate source	**A Technical**	
	1. Appearance	0
	2. Speed	3
	B Usability	
	1. Language	0
	2. Connection to task	0
	3. Audience	0
	4. Topicality	0
	C Verifiability	
	1. Reputation	0
	D. Reliability	
	1. Kind (site/PDF)	1
	2. Primary/secondary	0

Table 10

Means and Standard Deviations of Criteria Use

Sub skill	Criteria	All (N=46) Mean (SD)	Geography (N=16) Mean (SD)	Physics (N=16) Mean (SD)	Language and culture (N = 14) Mean (SD)
Evaluate search results					
	Title/Summary	4.7 (5.1)	5.7 (5.4)	4.2 (4.6)	4.2 (5.5)
	Rank in hit list	0.1 (0.2)	0.2 (0.4)	0.0 (0.0)	0.0 (0.0)
	Known to user	0.1 (0.3)	0.1 (0.3)	0.1 (0.3)	0.0 (0.0)
	Language	0.0 (0.2)	0.0 (0.0)	0.1 (0.3)	0.0 (0.0)
Evaluate information					
	Language	0.2 (0.6)	0.2 (0.4)	0.3 (0.8)	0.1 (0.5)
	Connection to task	1.9 (2.0)	2.4 (2.7)	2.1 (1.8)	1.2 (1.2)
	Amount	0.2 (0.4)	0.3 (0.4)	0.1 (0.3)	0.1 (0.3)
	Information agrees with prior knowledge	0.0 (0.2)	0.1 (0.3)	0.1 (0.3)	0.0 (0.0)
	Organization	0.0 (0.1)	0.0 (0.0)	0.1 (0.3)	0.0 (0.0)
	Kind of information	0.2 (0.7)	0.1 (0.5)	0.1 (0.5)	0.3 (1.1)
Evaluate source					
	Speed	0.1 (0.2)	0.1 (0.3)	0.0 (0.0)	0.1 (0.4)
	Kind	0.0 (0.1)	0.0 (0.0)	0.1 (0.3)	0.0 (0.0)

Table 11

Means and Standard Deviations of the Product Scores (Maximum = 9)

Tasks	N	Total Product Mean (SD)	Using more than one site (maximum = 1) Mean (SD)	Quality of Sources (maximum =4) Mean(SD)	Suitability of Information (maximum = 4) Mean(SD)
All	46	4.1 (1.7)	0.8 (0.4)	1.8 (0.8)	1.5(1.1)
Geography	16	3.6 (1.4)	0.8 (0.4)	1.9 (0.9)	0.9 (0.6)
Physics	16	5.1 (1.8)	0.8 (0.4)	2.0 (0.8)	2.3 (1.2)
Language and culture	14	3.5 (1.3)	0.9 (0.4)	1.6 (0.6)	1.1 (0.9)

Table 12

Criteria mentioned without prompting during group interviews

Sub skill	Criteria	Number of groups (maximum = 8)
Evaluate search results	1.Title/Summary	5
	2.Kind (site/PDF/)	0
	3.Address	2
	4.Rank in hit list	5
	5.Known to user	3
	6.Language	0
Evaluate information	**A Usability**	
	1. Language	5
	2. Connection to task	8
	3. Audience	2
	4. Topicality	0
	5. Amount	4
	B Verifiability	
	1. Author	4
	2. References	1
	3. Information agrees with more sites	8
	4. Information agrees with prior knowledge	1
	5. Organization	2
	C. Reliability	
	1.Kind of information	4
	2. Objectivity	0
	3. Primary/Secondary	0
	4. Goal	2
Evaluate source	**A Technical**	
	1. Appearance	6
	2. Speed	0
	B Usability	
	1. Language	0
	2. Connection to task	0
	3. Audience	0
	4. Topicality	0
	C Verifiability	
	1. Reputation	4
	D. Reliability	
	1. Kind (site/PDF)	2
	2. Primary/secondary	0

Figure 1. Time (%) Spent per Constituent Skill