Searchers' Relevance Judgments and Criteria in Evaluating Web Pages in a Learning Style Perspective

Chariste Papaeconomou
Royal School of Library and Information Science,
Birketinget 6
DK 2300 Copenhagen S, Denmark
chariste_p@yahoo.gr

Annemarie F. Zijlema
Royal School of Library and Information Science,
Birketinget 6
DK 2300 Copenhagen S, Denmark
afzijlema@hotmail.com

Peter Ingwersen*
Royal School of Library and Information Science,
Birketinget 6
DK 2300 Copenhagen S, Denmark
pi@db.dk

ABSTRACT

The paper presents the results of a case study of searcher's relevance criteria used for assessments of Web pages in a perspective of learning style. 15 test persons participated in the experiments based on two simulated work tasks that provided cover stories to trigger their information needs. Two learning styles were examined: Global and Sequential learners. The study applied eye-tracking for the observation of relevance hot spots on Web pages, learning style index analysis and post-search interviews to gain more in-depth information on relevance behavior.

Findings reveal that with respect to use of graded relevance scores and number of relevance criteria applied per task and test person there are no significant difference between the different styles. Although there differences are detected in the use of relevance criteria between Global and Sequential learners during assessments, they are statistically insignificant. When interviewed in retrospective the resulting profiles tend to become even similar across learning styles but a shift occurs from instant assessments with content features of web pages replacing topicality judgments as predominant relevance criteria..

Categories and Subject Descriptors

H.3 INFORMATION STORAGE AND RETRIEVAL

Conference Themes

Task-based interactive information retrieval and seeking behavior; Nature of relevance in contexts. Case study.

Keywords

Interactive information retrieval; Relevance assessments; Relevance criteria; Learning styles.

1. INTRODUCTION

Research has demonstrated that relevance is a concept that is personal [4], multidimensional, dynamic, and complex but measurable [25]. There is more to relevance than just the match between query and document.

This has resulted in altered ways of accomplishing empirical research in the area of IR system evaluation, e.g., as proposed by Borlund [5] where the information need, or the underlying work task, rather than the query is taken into consideration.

At the same time, research has been carried out on interface preferences of different cognitive style groups which helped them to accomplish their tasks. For example, the perception of multimedia by different cognitive styles [15] and personalization of Web services as digital libraries [14].

The present study is founded on the integrated cognitive view of information retrieval and information seeking advocated by Ingwersen & Järvelin [16] by investigating the contextual properties in Web documents of relevance assessments made by the information seeking actors. It aims to explore which aspects of Web pages the participants of different learning styles apply when making evaluation judgments. An empirical study was done with 15 test persons, who made relevance judgments on Web pages retrieved from two given simulated work tasks [5]. Especially, the experiments attempted to identify if any connections can be observed between the searchers' learning style and their relevance assessments, that is, if different learning styles lead to the use of different profiles of relevance criteria, e.g., in connection to graded relevance. The mentioned styles refer to the Sequential and Global learners.

Results of relevance studies based on learning styles are valuable since *if* they reveal patterns associated with the different styles that are discernable from each other, they may form a user model and ease personalized IR. Results thus serve as handles for search engine and Web designers as well as for the design of relevance feedback algorithms in interactive IR systems, matching particular user groups to relevant results.

Relevance has been studied a lot but thus far rarely with Web pages. A study on searchers' criteria for assessing Web pages by Tombros, Ruthven & Jose [28] showed that most criteria are common with paper based criteria, but that there are some additional criteria that should be taken into consideration. The present study moves one step further by using a different method and by taking learning styles into consideration.

IIiX'08, Information Interaction in Context, 2008, London, UK.

* Author to whom all correspondence should be addressed.

The paper is organized as follows. First, learning styles are briefly defined and research related to IR discussed. This is followed by a brief outline of relevance criteria investigations pertinent to the present experiments. The next section describes the methodological setup, including the use of simulated work task situations, learning style index examination, eye-tracking, explicit relevance assessments and post-interviews on relevance criteria used during the assessment processes. Section 4 demonstrates and discusses the results of the investigations, followed by concluding remarks.

2. LEARNING AND COGNITIVE STYLES

The learning style of a human actor is considered to be a fundamental element of interpreting information and modulating human cognition. When the searcher acquires information then hers/his current state of knowledge is being transformed into a new state, as cognition, which leads to knowledge and sometimes into making decisions. Our interest on learning style is justified in the sense that the style assumingly has implications on how a document's context is being interpreted and how the user's state of knowledge is being transformed [16, p. 29-30]. We consider searchers' relevance criteria as cognitive structures or cognitive representations of a Web document's context. When a searcher's relevance criteria (of a specific learning style) indicate the relevance of an information object, then we assume that people who have the same style might share the same criteria profile and more or less the same relevance assessments.

Learning style has been defined in several ways, e.g. by Valley as "[the] preference that an individual may have for processing information in a particular way when carrying out a learning activity" [30, p. 43]. This definition is quite similar to the explanations of *cognitive styles* by Messick [20, p.143]: "[Consistent] individual differences in preferred ways of organizing and processing information". For Riding and Rayner [24, p.7] it is "[an] in-built and automatic way of responding to information and situations…present at birth or at any rate is fixed early on in life and is thought to be deeply pervasive". In general, a cognitive style is a consistent tendency of how one for instance perceives, thinks or solves problems. Palmquist and Kim [21, p. 559] describes cognitive styles as "[exhibited] preferred modes or strategies that could be detected as distinctive or characteristic methods of performing".

Some semantic confusion occurs in the past years. However, we consider Rayner's point of view most appropriate [23, p.117] that cognitive and learning style are not exactly the same. For him "[if] learning style is carefully presented as a profile of the individual's approach to learning, cognitive style may be construed as part of a larger construct that might more properly be labeled a person's learning style." A person's learning style is made up of a core cognitive style which in turn influences learning processes and learning strategies. Similarly, we consider a person's cognitive style to influence activities like personal seeking processes or habits, in line with Byström & Järvelin [7].

There exist several learning style models. We have applied the Felder-Spurlin Model [11] with four distinctive types: *Global* vs. *Sequential* learners; and *Visual* vs. *Verbal* learners. Global learners tend to learn from large jumps, absorbing material almost randomly without seeing connections – and then suddenly 'getting it'. Sequential learners tend to gain understanding in logical linear steps, while Visual learners remember best what they see, pictures, diagrams, time lines, whilst Verbal learners prefer words and written or spoken explanations.

2.1 Related Studies on Learning Styles in Information Retrieval

Liu & Reed [19] found that hypermedia was explored both in a non-linear (Global) mode and sequentially. Leader & Klein [18] observed that 'field independent' searchers (like Global learners) obtained better retrieval performance than Sequential ones, especially when the system was provided with non-linear navigation and analytic searching. Palmquist & Kim [21]found that the two groupings, when experienced, were spending the same amount of search time and clicks. Less experienced Sequential-like searchers, however, did require more time and clicks during their search.

With respect to Verbal vs. Visual learner groups Ford et al. [13] observed an association between poor retrieval and the Verbal style. Also, they found that searching effectiveness was linked to the Visual style, males and low cognitive complexity. Their outcome on Internet perception concurs with a previous study by Ford & Miller [12]. In an earlier study on hypertext navigation Ellis et al. [9] detected that serialists (like Sequential learners) made greater use of a keyword index.

2.2 Pertinent Relevance Criteria Studies

Barry [1] studied the users' criteria when evaluating the information within documents. In this study, "[relevance] was conceptualized as any connection that existed between the user's information need situations and the information provided by the documents" (p. 152). This type of relevance has been called 'Intellectual Topicality' [4]. Participants were asked to examine the given documents and *circle* every portion of the document that had "prompted a reaction to pursue some aspect of the document" (p. 153). This was followed up by an interview with the subject to discuss the portions that the subject had circled. Barry's study resulted in 23 relevance criteria grouped in 7 relevance categories. The four most important groups of criteria she revealed were pertaining to "information content of the document" (e.g. topic, depth/scope), "to user's background and experience" (i.e. ability to understand, novelty), "users belief and preferences" (i.e. subjective accuracy, effectiveness) and "to other information and sources within the environment" (i.e. consensus, availability within the environment) (p.157). Later Barry and Schamber [2] developed a quite distinctive categorization of relevance criteria, building on several large-scale empirical relevance studies.

Tombros et al. [28] investigated the criteria of users when judging Web pages. 24 test persons searched the Web to fulfill 3 search tasks, which were embedded in one simulated work task. They were free to search on the Web and to use the search engines they wanted to use. They were asked to think aloud during searching, to mention the *features* that helped them to assess. The most important features were grouped in 5 categories (p. 2): 1. *Text*: Content, Numbers, Titles/Headings, Query Terms, Text Quantity; 2. *Structure*: Layout, Links, Links Quality, Table Layout; 3. *Quality*: Scope/Depth, Authority/Source, Currency, General Quality, Content Novelty; 4. *Non-textual*: Pictures; and 5. *Physical Properties*: Page Not Found. Page Location, Page Already Seen, Others.

In our investigation, we are pursuing the methodological procedures and criteria by Barry [1] and Barry & Schamber [2] as

well as most of the criteria categories observed by Tombros et al. [28]. With respect to the application of eye-tracking we use this device as a *tool* for obtaining deeper information on relevance criteria associated with assessments and learning styles, replacing Barry's manual relevance circling method applied to judged documents, mentioned above. Other eye-tracking experiments with reference to relevance feedback purposes in IR are referred to in the methodological section 3.2.3.

3. METHODOLOGICAL SETUP

15 test persons participated in the experiments which took place in the IT Laboratory of the Royal School of LIS, Copenhagen. Eight persons were women and seven were men. Most of the participants (n=10) had an educational background in LIS on bachelor or MSc levels. The remaining (n=5) came from various disciplines, like sociology, computer science, biology and chemical engineering. 27 % considered themselves as average users of the Web (i.e. knowing how the Web is structured but cannot use it optimally) and 73 % as familiar with it (having a more in-depth knowledge of the Web).

3.1 Experimental Procedure

For each participant an individual lab session was arranged. The session was started by introducing to the participant the whole procedure. It contained four parts. First, the test person would be placed at the eye-tracker computer screen. Eye calibration would take place and explained. How to navigate the test system was also demonstrated for the test person, who could try out the procedure. Then he/she received the first of two cover stories (simulated work task situations) to be read. Six Web pages retrieved beforehand by the research team and shown on the screen was then to be assessed one by one for relevance (11-point scale) according to the participant's interpretation of the cover story. For each page the activity time was set to 4 minutes. Participants were allowed browsing the Web initiated by the Web page in question, but it was repeatedly emphasized that they should assess the provided *Web page*, based on its features, not the entire site. This instant assessments for each page were done on a open-ended paper form prior to seeing the next Web page. Test persons were allowed to return to a page and re-assess it.

Secondly, after navigating and assessing all six given Web pages belonging to the first simulated task a post-interview took place as retrospective think aloud. The test person was asked to talk about what he/she was doing when reading something on the Web page in question, how they felt about that page and the relevance criteria applied. The interview was based on the gaze recordings of the eye tracker that were replayed for the person. In contrast to common screen capture eye-tracking recordings make available the points of gazing on the screen. The interview was audio recorded with consent of the participant and recordings turned into verbal protocols for later analysis with respect to categorization and coding of relevance criteria. The reason for the post-interview was to obtain supplementary information on the assessment process and criteria. An additional reason was that although eye tracking reveals that a person gazes at something in the document, it is unknown if the person actually is reading or take the spot into consideration as a criterion.

Third, the participant repeated the procedural steps 1-2 on the second simulated works task situation provided by the research group.

Fourth, each participant ended the investigation by filling out two questionnaires. One concerned demographic and alike data; the other was the test which revealed learning style – the so-called Index of Learning Styles (ILS) [11].

3.2 Experimental Devices

The relevance criteria categories used by Barry & Schamber [2] and by Tombros et al. [28] were initially used as coding schemes in the verbal protocols and written relevance assessment forms. After some re-coding of data from the latter form and the response data recorded during the post interview the final scheme ended up in 12 major categories, with the central 'Web page content' category containing 8 sub-categories – see e.g. Table 2 below. The consistency of the final scheme was tested by comparing the coding of selected verbal protocols across the individual research team members.

3.2.1 Simulated Work Task Situations

In order to assure experimental control *and* realism the same two simulative work task situations were assigned each participant, as proposed and verified by Borlund [3; 5]. A simulated work task situation is seen as a cognitive state, which creates an information need that has to be satisfied so that the user is able to deal with the situation and move on. The realism by using the situation comes from the involvement of the tests person who, based on the work task assigned, develops his/her *own* individual and subjective information need. The participants then dynamically assessed relevance of the retrieved objects, in relation to their own interpretation of the simulated work task, as in real life [3, p.77]. Specifically the simulated work task situation helps to describe: i) the source of the information need; ii) the context of the situation, i.e., the problem which has to be solved; and iii) serves to make the participant understand the objective of the search [6, p.229]. It is thus much more encompassing than an assigned topical request as commonly applied in TREC.

In our experimental setting both simulated task situations were factual and semantically open. One concerned the "Olympic Games in Beijing", the other the "Riverdance Show":

> Beijing is hosting in 2008 (8th-24th August) the Olympic Games. A friend of yours, who is a big fan of the Olympic Games, wants to attend the events and asks you to join in this trip. You find this invitation interesting. You are not a big fan of the games but you always wanted to visit China, therefore you want to find information about the sightseeing in the city and the activities that the Chinese will offer during the games. Find for instance places you could visit, activities you could do in relation to the Chinese culture or in the spirit of the games.
>
> Last month a friend of yours watched the live show of Riverdance performed by traditional Irish step dancers and he thought that is was amazing. You had never heard about it before and you wonder what it is and what the origins of this dance are.

The degree of semantic openness determines the range of a test person's interpretations. In both simulated work tasks, a variety of interpretations existed which reveal, in our opinion, the individuals' plurality of different cognitive states during the information processing procedure.

For each simulated task the research team (2 persons) selected 6 pages among the top-20 retrieved pages pooled from several searches on Google. First, the pooled pages were assessed for relevance by the two research team members applying the same 11-point scale as the test persons. Only Web pages with

agreement upon the assessments within 3 scaling points could be selected. Web pages were seen as 'relevant' (score 7-10), 'partially relevant (score 3-6) and 'not relevant' (score 0-2). Secondly, this produced a set of Web pages per simulated task, from which the six final Web pages were selected for each work task. Third, this was done according to five common criteria (the order does not indicate any significance): 1. *Web structure and layout*: Are there indexes, files, bookmarks, internal links/ external links, short/ long page, pictures, photos, videos, commercials, clear paragraphs? How deep does the test person have to navigate? Is the Web page a kind of specific type like an online article? 2. *Web genre*: Is the page part of an official, commercial, or personal Website? 3. *Socio-cognitive relevance* [8]: Level of relevance according to the researchers as agreed above. 4. *Time of creation or update*: Is the information old? Or how current is the Web page/ Website? 5. *Content*: Does it cover the simulative work task situation? Is it easy to read or interesting?

For each simulated task one Web page was chosen as considered 'non relevant' to the task by the research team members, i.e., assessed within the the scoring range 0-2. It should be noted that both team members' learning styles were Sequential. In the experimental setting both the simulated tasks and their Web pages were distributed to the test persons according to Latin Square principles, in order to avoid learning effects. Each of the 15 test persons would thus assess both tasks and 12 Web pages; in total 180 Web page assessments were collected.

3.2.2 The Index of Learning Style Extraction

In order to extract the participants' learning style the Index of Learning Style (ILS) was used in a paper pencil version, and the answers were later transferred into the on-line version. The ILS is a forty four question instrument designed to assess the preferences in the Felder-Silverman model, developed at North Carolina State University. It is considered to be reliable since the scores of test-retest reliability measurements are satisfactory according to Felder & Spurlin [11, p. 107]. According to these measurements, it is proved that the ILS reflects a preference or attitude. There is a group of eleven questions for each of four types grouped in two dimensions: Global vs. Sequential and Verbal vs. Visual. They result in a 'style' for each person located on a scale between +/- 11, e.g. from 'highly Global' over 'Global/Sequential' to 'highly Sequential'.

3.2.3 Eye-Tracking

Eye-tracking has been applied to relevance and IR research on several occasions, foremost as an interface device for obtaining implicit relevance information for relevance feedback purposes. Puolamäki et al. [22] studied proactive IR by combining implicit relevance feedback and collaborative filtering. The implicit feedback was inferred from the eye movement signals (by using the Tobii 1750 eye tracker). Their task was to predict relevance and to study how far they could go by measuring implicit feedback signals from the user and combining them with existing data on preferences of similar-minded groups. The key assumption that motivated them to use eye movements is that "attention patterns correlate with relevance, and that attention patterns are reflected in eye movements" (p. 146-147). More over, Salojärvi et al. [26] showed that relevance can be inferred from eye movements, at least to a certain degree.

Also in 2005 Joachims et al. [17] examined the reliability of implicit feedback generated from click through data in a WWW search. They analyzed the user's decision process by using the eye tracking and comparing implicit feedback against manual relevance judgments. They used the eye tracker in order to understand how users behave on Google's results page, how users interact with the list of ranked results and how their behavior can be interpreted as relevance judgments. The use of the eye tracking provided the researchers with a detailed insight into the user's decision making process. Their results showed that "[users] make informed decisions among the abstracts they observe and that clicks reflect relevance judgments" (p.154).

In common to these and other IR-related eye-tracking experiments the test persons involved were always seen as alike with common characteristics. Our experiments attempt to observe if different learning characteristics influence the relevance assessments.

In our tests we also applied the Tobii 1750 Eye Tracker as a *tool for observation*, not as an integrated part of relevance feedback experiments. It is a remote near-infrared tracker with a sampling rate of 50 Hz and accuracy of one degree [10]. By showing the objective eye tracking recording to the participant his/her interpretations of what occurred on the screen and which criteria that were used helped to further the understanding of the decision process. A second aim was to use the Clear View analysis software provided by Tobii to visualize the *hot spot plots*, which are 'photos' of the test pages: One may observe where a given group of the participants commonly were gazing. We managed to obtain 3 test Web pages demonstrating reliable *generalized plots* of the different learning style groupings, one of which is demonstrated as Appendices 1-2 for Global learners and Sequentials.

4. EXPERIMENTAL RESULTS

The learning style tests resulted in a breakdown into the four style groups of the 15 test persons. However, the style groups were not equally distributed among the test persons.

Global (n=10) and Sequential learners (n=5) were chosen for further analyses, whilst Visuals (n=11) were discharged since 7 also were Global learners. Verbals only counted for four participants out of which 3 also were Sequential learners. These reduced analyses were obviously caused by the application of the ILS index as a post-screening device. Time constraints did not allow for additional adequate participants to be included in the study.

4.1 Graded Relevance Assessments and Learning Style

Each test person assessed the relevance of each assigned Web page per task by means of an 11-point scale, with value 0 (zero) as 'totally non-relevant' and value 10 as 'highly relevant'. The scale was used in a gliding manner. Table 1 shows the grading per Web page for both tasks for the two styles. One should note that the Web pages assessed by the two research team members on the tables are coded 'Not Relevant' (score range 0-2), 'Partial Relevant' (range 3-6), and 'Relevant' (range 7-10).

Table 1 demonstrates that there are some but not significant differences between the two learning styles' grading of relevance assessments across the two search tasks – on average 6.1 and 4.5, respectively.

The observed differences between the graded relevance assessments adhere primarily to the nature of the two different work task simulations. The Riverdance Web pages are regarded less relevant than the Beijing pages. With respect to the Beijing task the Sequentials are the only ones agreeing with the research team on the non-relevance of Web page A (score 1.6), which also lower their average score. Without that score the averages for the two styles are more similar. As regards the Riverdance task, on average all test persons over both styles agree on Web page A as 'non relevant'. One observes some very high and low standard deviation (SD) scores – page A and F in the Beijing case and page A and C(D) in the Riverdance task. In general, the Sequential learners judge the Beijing task less relevant over 5 of 6 pages, while the opposite is the case in the Riverdance task.

Aside from the Beijing task page A, the correlation between the research team evaluations of the pages and those of the test persons is high. In particular, all partially relevant pages (as assessed by the research team) are also judged partially relevant by the test persons (score range 3-6) – regardless learning style.

Table 1. Average graded relevance scores of six Web pages (A-F) per search task over two learning styles and 15 test persons. SD: Standard Deviation.

Beijing Task	***Not Rel.*** *Page A*	**Relevant** Page B	**Relevant** Page C	Partial Rel. Page D	Partial Rel. Page E	**Relevant** Page F	Average
Globals (10)	5.5	8	7.4	6.8	4.7	6.7	*6.5*
Sequentials (5)	1.6	7.4	6.1	4.6	3.4	8.4	*5.2*
Both styles, mean	4.2	7.8	6.9	6.1	4.3	7.3	*6.1*
Max. SD	**2.6**	2.5	2.5	2.6	2.6	**1.4**	

Riverdance Task	***Not Rel.*** *Page A*	Partial Rel. Page B	Partial Rel. Page C	**Relevant** Page D	Partial Rel. Page E	**Relevant** Page F	Average
Globals (10)	0.4	5.4	3.6	8.6	3.1	5.8	*4.5*
Sequentials (5)	0	3.6	4.2	9.4	4.6	5.6	*4.6*
Both styles, mean	0.3	4.8	**3.8**	8.9	3.6	5.7	*4.5*
Max. SD	**1.3**	2.7	**4.3**	1	2.6	3.8	

These findings indicate that learning style seems to have only a *minor effect* on searchers' relevance assessment scores when graded scales are applied in experiments. But the sample size is too small to provide more significant conclusions.

4.2 Relevance Criteria Applied to Instant Assessments of Web Pages

Table 2 displays the application of criteria when test persons were assessing the Web pages directly one by one with the assessment statements written down on an open-ended form. Bold figures signify top-5 criteria; figures in italics mean very insignificant or no application of a criterion.

First, we observe that regardless the learning style the number of relevance criteria applied during instant assessments per task and person remains almost the same over the six Web pages (12.7-13.85), or 2.1 vs. 2.3 criteria used per page. No steady pattern could be observed as to applied combinations of criteria.

Table 2. Use of relevance criteria during instant relevance assessments of Web pages related to both work task situations and over 15 test persons

Instant assessments	**Globals**		**Sequent.**		**Task 1+2**
Relevance Criteria used	*(n=10)*	Percent	*(n=5)*	Percent	*Profile %*
Depth/scope specificity	75	**27.08**	29	**22.83**	**26**
Content of web page		**25.63**		**25.99**	**25.75**
Table of contents/sitemap	1	*0.36*	0	*0.00*	0.25
Title, headline, captions	3	*1.08*	0	*0.00*	0.74
Keywords	1	*0.36*	0	*0.00*	0.25
Link anchor text	29	**10.47**	20	**15.75**	12.13
Multimedia	19	**6.86**	9	**7.09**	6.93
Recommendations	6	2.17	1	*0.79*	1.73
Advertisements	4	1.44	1	*0.79*	1.24
Other content aspects	8	2.89	2	1.57	2.48
Topic of web page/text	67	**24.19**	43	**33.86**	**27.23**
Accuracy/validity	10	3.61	0	*0.00*	2.48
Clarity	6	2.17	0	*0.00*	1.49
Currency	4	1.44	5	3.94	2.23
Accessibility	4	1.44	2	1.57	1.49
Affectiveness	6	2.17	3	2.36	2.23
Web page layout	24	**8.66**	7	**5.51**	**7.67**
Pers. background knowledge	1	*0.36*	0	*0.00*	0.25
Target group/purpose of page	9	3.25	5	3.94	3.47
Content novelty	0	*0.00*	0	*0.00*	0
Total	277	100.00	127	100	100
Mean no. of criteria per person	13.85		12.7		

Secondly, some criteria are almost not mentioned by the test persons, like 'Table of Contents' and 'Keywords' or 'Personal Background Knowledge' (percent in *italics*), regardless style.

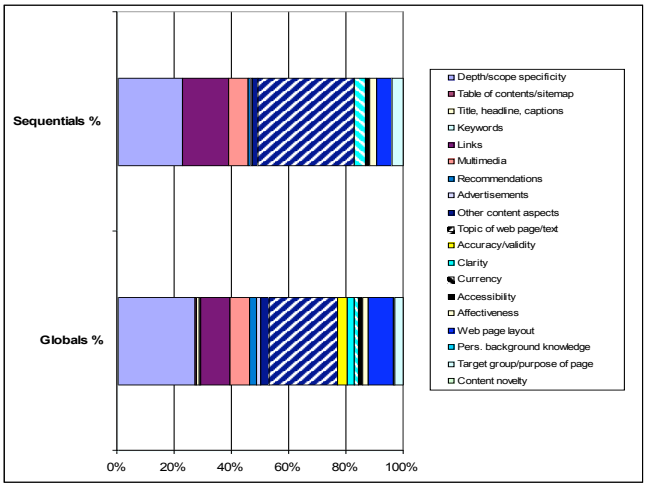


Diagram 1. Relevance criteria profiles for Global and Sequential learners at instant assessments of Web pages from two simulated tasks over 15 test persons.

Third, at these instant screen assessments some criteria are preferred to a larger extent in one style than in others: 'Depth/Scope' (27%) and 'Web page layout' (8%) among Globals and 'Topic of the Web page' (almost 34%) and the sub-category criterion 'Link anchor text' (16%) among Sequential learners. Thus, one observes a noticeable difference between Global and Sequential learners. However, a correlation test (Pearson) of the

two profiles reveals that r = .944 (p=.01; CV = .575; DF = 17), implying a high correlation over the 19 pairs. The two profiles are consequently *not significantly different* from one another.

Diagram 1 illustrates the two profiles from the Global and the Sequential learners based on the Table 2 data. The sequence of colors from left to right follows the criteria sequence in Table 2.

4.3 Relevance Criteria Applied during Retrospective Eye-Tracking Interviews

Table 3 demonstrates that almost equivalent patterns appear across the Global and Sequential learners – with a dominance of the '*Content of Web page*' category applied by Sequentials. Figures in bold and italics are as in Table 2.

Table 3. Use of relevance criteria during retrospective relevance assessments of Web pages related to both work task situations and over 15 test persons

Retrospective assessments	**Globals**		**Sequent.**		**Task 1+2**
Relevance Criteria	*(n=10)*	Percent	*(n=5)*	Percent	Profile %
Depth/scope specificity	79	**13.17**	41	**13.85**	**13.39**
Content of web page		**39.83**		**44.59**	**41.41**
Table of contents/sitemap	4	*0.67*	2	*0.68*	*0.67*
Title, headline, captions	39	6.50	22	**7.43**	6.81
Keywords	22	3.67	14	4.73	4.02
Links	76	**12.67**	47	**15.88**	**13.73**
Multimedia	51	**8.50**	27	**9.12**	**8.71**
Recommendations	9	1.50	5	1.69	1.56
Advertisements	14	2.33	3	1.01	1.90
Other content aspects	24	4.00	12	4.05	4.02
Topic of web page/text	80	**13.33**	44	**14.86**	**13.84**
Accuracy/validity	48	8.00	14	4.73	6.92
Clarity	24	4.00	8	2.70	3.57
Currency	7	1.17	5	1.69	1.34
Accessibility	14	2.33	6	2.03	2.23
Affectiveness	31	5.17	11	3.72	4.69
Web page layout	46	**7.67**	19	**6.42**	**7.25**
Pers. background knowledge	9	1.50	5	1.69	1.56
Target group/purpose of page	19	3.17	11	3.72	3.35
Content novelty	4	*0.67*	0	0.00	*0.45*
Total	600	100.00	296	100	100.00
Mean no. of criteria per person	30		29.6		

However, exactly concerning the 'Content of Web page' category a *marked shift* occurs from Table 2 to Table 3: Over all learning styles its percentage increases from approx. 25 % to above 40 % - at a cost of the 'Topic of Web page' criterion, which served as the dominant relevance criterion together with 'Depth/scope'.. *Content features* become thus that *context* most used for relevance assessments. This concerns both learning styles. The correlation between the two profiles is very high. They are highly similar in a statistical sense.

In addition one observes that during the retrospective interviews and assessments the average number of criteria applied increases to 30 and 29.6 for Global and Sequential learners, respectively. These numbers correspond to 5 different relevance criteria applied per Web page. No difference between the learning styles can be detected.

In line with Table 1, Table 4 further demonstrates that the number of criteria used depends on the simulated work tasks assigned the test persons rather than the learning styles. In Table 4 the use of bold and italics corresponds to that in Tables 2-3. Table 4 also demonstrates that the Sequential learners to a higher degree replace topicality by content features than Global learners, compared to Table 2. 'Keywords' and 'Title, headings, captions' (in the Riverdance task) and 'Multimedia' become top relevance criteria for both learning styles. The effect of the assigned simulated work tasks on the patterns of criteria used is noticeable.

Table 4. Use of relevance criteria per task during retrospective assessments of Web pages over 15 test persons.

	Beijing				**Riverdance**			
Retrospective assessments	Globals		Sequent.		Globals		Sequent.	
Relevance Criteria used	*(n=10)*	Percent	*(n=5)*	Percent	(n=10)	Percent	(n=5)	Percent
Depth/scope specificity	44	**13.13**	23	**13.609**	35	**13.21**	18	**14.17**
Content of web page		**37.91**		**42.60**		**42.26**		**47.24**
Table of contents/sitemap	2	*0.60*	2	1.18	2	*0.75*	0	0.00
Title, headline, captions	20	5.97	12	**7.10**	19	7.17	10	**7.87**
Keywords	5	1.49	3	1.78	17	6.42	11	**8.66**
Links	44	**13.13**	27	**15.98**	32	**12.08**	20	**15.75**
Multimedia	24	7.16	15	**8.88**	27	**10.19**	12	**9.45**
Recommendations	9	2.69	5	2.96	0	*0.00*	0	0.00
Advertisements	8	2.39	1	*0.59*	6	2.26	2	1.57
Other content aspects	15	4.48	7	4.14	9	3.40	5	3.94
Topic of web page/text	36	**10.75**	20	**11.83**	44	**16.60**	24	**18.90**
Accuracy/validity	26	**7.76**	8	4.73	22	**8.30**	6	4.72
Clarity	13	3.88	6	3.55	11	4.15	2	1.57
Currency	7	2.09	5	2.96	0	*0.00*	0	0.00
Accessibility	13	3.88	5	2.96	1	*0.38*	1	0.79
Affectiveness	21	6.27	8	4.73	10	3.77	3	2.36
Web page layout	33	**9.85**	12	7.10	13	4.91	7	5.51
Pers. background knowledge	7	2.09	4	2.37	2	*0.75*	1	0.79
Target group/purpose of page	5	1.49	6	3.55	14	5.28	5	3.94
Content novelty	3	*0.90*	0	*0.00*	1	*0.38*	0	0.00
Total	335	100	169	100	265	100	127	100
Mean no. of criteria per person	33.5		33.8		26.5		25.4	

Tables 3-4 also show that almost all the criteria become applied during the retrospective thinking aloud of assessments triggered by the in-depth post-interviews. The patterns for the different styles are quite consistent with the overall profile, Table 3.

5. DISCUSSION

Regardless the data collection method (at assessment time or later with eye-tracking replay and retrospective thinking aloud during post interviewing) we observed that the number of criteria used per person per task was conceivably similar across the different learning styles. This similarity also concerned the relevance scores used. It seems to be the *nature of the work task* that affects the number of criteria and relevance scoring used rather than style. It would hence be recommendable, and in line with Borlund [3; 5] and other methodological discussions [16], to apply more than two simulated tasks, probably 6-8.

Only at the time of instant assessment of web pages the applied relevance criteria demonstrated patterns distinguishing Global learners from Sequentials. The former applied 'Depth/Scope' and 'Web layout' as major criteria whilst Sequentials depended strongly on 'Link anchor text' and 'Topic of Web page' criteria. However, the detected differences in relevance criteria profiles were not statistically significant.

Appendices 1-2 demonstrate 'hot spot' pictures of Global learners vs. Sequentials addressing the same Web page during instant assessments. Sequentials interact as expected by gazing from left to right following the layout of the page. This behavior probably leads to interpretations of the page's aboutness (topical relevance). In contrast, the Global learners applied more diffuse modes of gazing at the same page and it was not possible to create a common (numbered) hot spot pattern of gazing the page., Without doubt (groups of) people tend to read and interpret Web pages in different ways during instant relevance assessments, but there do not seem to be significant differences in their assessment behavior and scoring.

When in-depth interviews and retrospective thinking aloud were used for data collection almost all criteria were applied, doubling the number of criteria stated by the participants. This was anticipated but it was also believed that (different) learning style profiles detected at the earlier stage of the investigation would be repeated or even strengthened by the post-interview. Instead, the post-interviews triggered a *shift* in the learning style profiles, so that they became even more alike. The web page content features replaced topical assessments of pages as the dominant set of relevance criteria, across the two learning styles, but most radical for the Sequentials.

What seems to happen during the retrospective interview phase is that the test persons start to explain their 'topicality assessments' by invoking exactly those Web page features that triggered them to assess the aboutness of pages. Their statements during the interview and reply of eye-tracking become thus detailed explanations of former statements that now become coded differently. Hence the shift from 'topicality of web page' to 'content features'. These two criteria categories are highly related and the result mirrors the Vakkari findings on academic bibliographic records that the most employed relevance criterion associates to topicality, aboutness and contents [29].

Methodologically speaking the analyses demonstrate that the application of a detailed relevance criteria scheme is necessary for achieving any indicative results. This is in line with the Tombros et al. [28] study. On the other hand, the study also reveals that in the case of instant assessments one may capture approximately 2-3 different criteria per person per Web document judged. This number doubles when the much more cumbersome retrospective thinking aloud and interviewing takes place. As shown, the profiles of applied criteria somewhat change with in-depth interviews of assessment behavior – Table 2 vs. Table 3.

One should notice that if learning style was *not* applied as a variable during data capture and analysis, the 'criteria profile' from the investigation over six Web pages and both tasks and 15 test persons would look like the right-most column (Task 1+2 Profile), Tables 2-3. These profiles concurs with the findings by Tombros et al. [28] concerning central criteria concerned with Depth/Scope, Topic, Links, Page Layout and Content-associated Web page features, the latter over all styles: 25.75%, Table 2, increasing to 41.41%, Table 3.

Owing to the differences demonstrated between the two simulated work task situations in terms of number of relevance criteria used during assessments and the judged relevance scores, we recommend the use of more than two tasks, as stated above, regardless the number of test persons participating, if the study intends to investigate *retrieval performance* dealing with different searcher groups in a statistically significant way. This is in line with findings by Vorhees [31] concerning the TREC environment.

However, our investigation also showed that the number of participants should be increased (from 15) in order to study human information behavior, like relevance *assessment behavior*, e.g., to 20-30 test persons. Most importantly one should ensure that the number of *test persons* covers all the desired variables in a sufficient number, e.g., 2 x 10 in case of two groups, or 4 x 10 as minimum in the case of four different but associated groups, like four learning styles. Because the selection of participants was done without pre-screening their learning styles too few Sequential learners and none Verbal or Visual learners became included in our analyses.

6. CONCLUSION

In summary, the study should have included more test persons adhering to different learning styles in a substantial and preferably equal number. It would also have improved the validity of findings to have included more simulated work task situations to serve as cover stories for the relevance assessments. In particular, the significance of the 'non findings' in the present study are less convincing owing to the small sample size.

Given these methodological drawbacks, the present study demonstrates that the effect of learning styles (Global versus Sequential) on relevance assessments of Web pages seems rather small. In particular our findings indicate that relevance scores and the number of relevance criteria applied during judgments per person and Web page is constant across learning styles and primarily affected by the differences in the assigned work task situations. With retrospective thinking aloud and post-interviews more in-depth information on relevance criteria is indeed captured and a rather profound shift takes place between the application of topical relevance and content features as criteria, the latter gaining a strong momentum. In a relevance criteria perspective the findings support previous ones in relation to Web information retrieval.

7. REFERENCES

[1] Barry, C.L. 1994. User defined relevance criteria: an exploratory study. *JASIS*, 45(3), 149-159.

[2] Barry, C. & Schamber, L. 1998. User's criteria for relevance evaluation: A cross situational comparison. *Inf. Proc. & Man.*, 34(2/3), 219-236.

[3] Borlund, P. 2000. Experimental components for the evaluation of interactive information retrieval systems. *J. of Doc.*, 56 (1), 71-90

[4] Borlund, P. 2003a. The concept of relevance in IR. *JASIST*, 54(10): p. 913-925.

[5] Borlund, P. 2003b. The IIR evaluation model: A framework for evaluation of interactive information retrieval systems. *Inf. Research*, 8(3), paper no.152. Available: http://informationr.net/ir/8-3/paper152.html.

[6] Borlund, P. & Ingwersen, P. 1997. The development of a method for the evaluation of interactive information retrieval systems. *J. of Doc.*, (53)3, 225-250.

[7] Byström, K. & Järvelin K. 1995. Task complexity affects information seeking and use. *Inf. Proc. & Man., 31*(2): 191-213.

[8] Cosijn, E. & Ingwersen, P. 2000. Dimensions of relevance. *Inf. Proc. & Man*, 36, 533-550.

[9] Ellis, D., Ford, N., & Wood, F. 1992. *Hypertext and learning styles*. Final Report of a project funded by the Learning technology Unit. Sheffield: Employment Department.

[10] Ewing, K. 2005. *White paper: studying web pages using eye tracking*. Tobii Technology, August 2005, p15.

[11] Felder, R.M. & Spurlin, J. 2005. Applications, reliability and validity of the Index of Learning Styles. *Int. J. Eng. Edu.*, 21(1), 103-112.

[12] Ford, N., Miller, D. (1996). Gender differences in Internet perception and use. In: *Electronic Lib. & Vis. Inf. Res.* Papers

from the third ELVIRA conference, 30 April 1996, 87-202. London: ASLIB.

[13] Ford, N., Miller, D. & Moss, N. 2001. The role of individual differences in Internet searching: An empirical study. *JASIST,* 52(12), 1049-1066.

[14] Frias-Martinez, E., Chen, S.Y. & Liu, X. 2007. Automatic Cognitive Style Identification of Digital Library Users for Personalization. *JASIST*, 58(2): p. 237-251.

[15] Ghinea, G. & Chen, Y. 2003. The impact of cognitive styles on perceptual distributed multimedia quality. *British J. of Edu. Tech.*, 34 (4): p. 393-406.

[16] Ingwersen, P. & Järvelin, K. 2005. *The turn: integration of information seeking and retrieval in context*. Netherlands: Springer Verlag. xiv, 448p.

[17] Joachims, T., Granka, L., Pan, B., Hembrooke, H. & Gay, G. 2005. Accurately interpreting clickthrough data as implicit feedback. In: *SIGIR'05 Proc*, August 15-19, Salvador, Brazil, 154-161.

[18] Leader, L.F. & Klein, J.D. 1996. The effects of search tool type and cognitive style on performance during the hypermedia database searches. *Edu. Tech. Res. & Dev.*, 4(1), 24-51.

[19] Liu, M. & Reed, W.M. 1994. The relationship between the learning strategies and learning styles in hypermedia environment. *Comp. in Hum. Behav.*, 10(4), 419-434.

[20] Messick, S. 1984. The nature of cognitive styles: Problems and promise in educational practice. *Edu. Psych.*, 19(2), 59-74.

[21] Palmquist, R.A. & Kim, K.S. 2000. Cognitive style and on line database search experience as predictors of web search performance. *JASIST*, 51(6), 558-566.

[22] Puolamäki, K., Salojärvi, J. & Savia, E. 2005. Combining eye movements and collaborative filtering for proactive information retrieval. In: *SIGIR'05 Proc.*, August 15-19, Salvador, Brazil, 146-153.

[23] Rayner, S. G. 2000. Reconstructing style differences in thinking and learning: profiling learning performance. In: *Int. Perspec. on Indiv. Diff. Vol. 1 Cognitive Styles*, ed. by R. J. Riding & S. G. Rayner. Stamford, Connecticut: Ablex Publ. Corp., 115-177.

[24] Riding, R, J., Rayner, S.G. 1998. *Cognitive styles and learning strategies*. London: David Fulton.

[25] Schamber, L., Eisenberg, M.B., Nilan, M.S. 1990. A re-examination of relevance: Towards a dynamic, situational definition. *In. Proc. & Man.*, 26(6), 755-775.

[26] Salojärvi, J., Kojo, I., Simola J. & Kaski, S. 2003. Can relevance be inferred from eye movements in information retrieval? In: *Proc. of WSOM'03, Workshop on Self-Org. Maps:* 261-266. Kyushu Institute of Technology, Kitakyushu, Japan.

[27] Santally, M.I. & Senteni, A. 2005. A learning object approach to personalized web-based instruction. Available: http://www.eurodl.org/materials/contrib/2005/Santally.htm

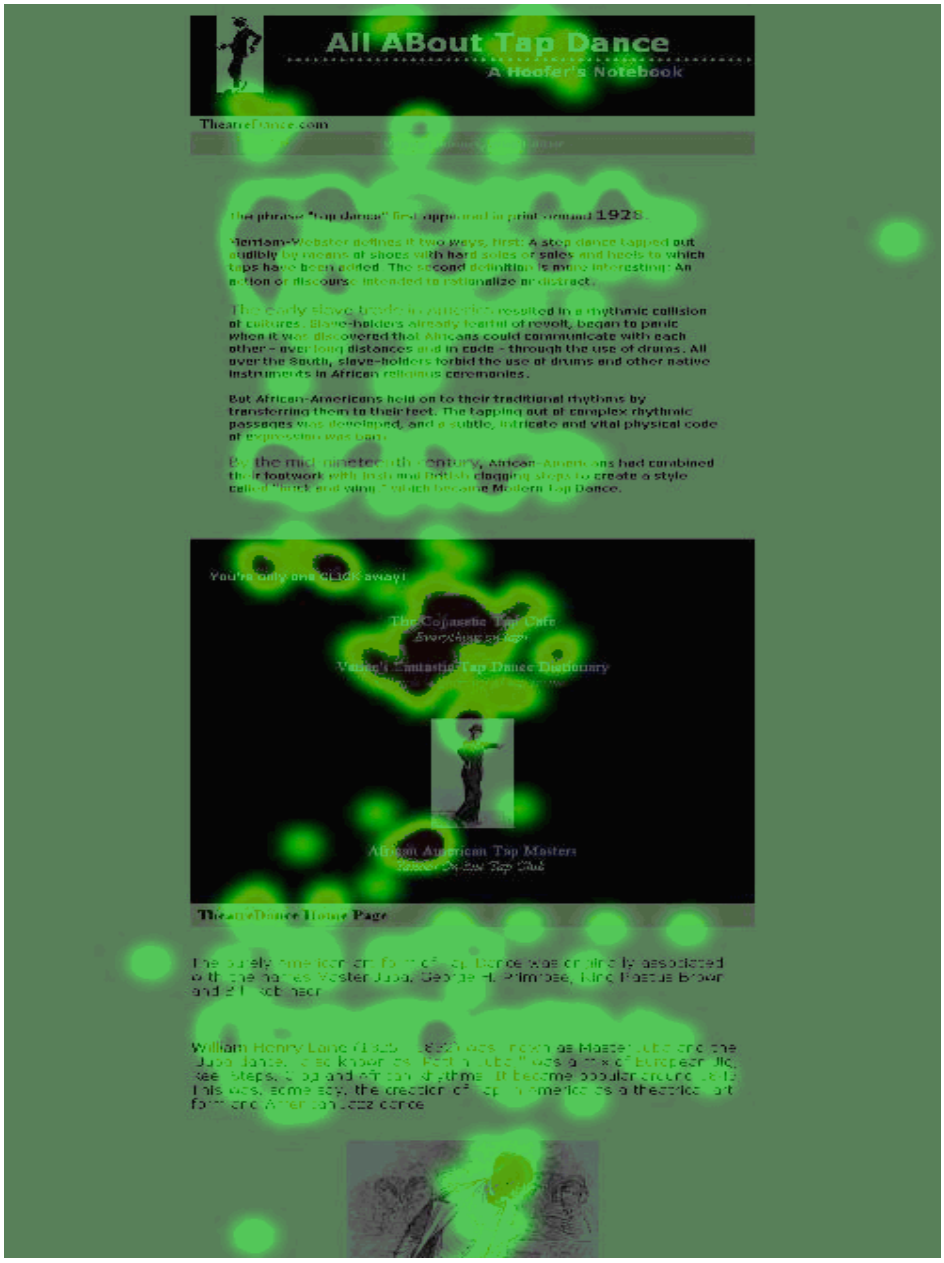
[28] Tombros, A., Ruthven, I., Jose, J.M. 2003. Searchers' criteria for assessing web pages. In:*SIGIR'03 Proc.*, July 28–August 1, 2003, Toronto, Canada.

[29] Vakkari, P. 2001. Changes in search tactics and relevance judgments in preparing a research proposal: A summary of findings of a longitudinal study. *Inf. Retrieval, 4*(3/4), 295-310.

[30] Valley, K. 1997. Learning styles and courseware design. *Ass. Learn. Tech. J.,* 5(2), 42-51.

[31] Voorhees, E.M. 1998. Variations in relevance judgments and the measurement of retrieval effectiveness. In: *SIGIR-98 Proc.*, 315-323.

Appendix 1: Global learners 'hot spot' from eye-tracking of page 3, Riverdance task.



Appendix 2: Sequentials 'hot spot' from eye-tracking of page 3, Riverdance task.