Characterizing Relevance with Eye-tracking Measures

Jacek Gwizdka
School of Information, University of Texas
1616 Guadalupe St., Austin, TX 78701, USA
iiix2014@gwizdka.com

ABSTRACT
Relevance, a fundamental concept in information search and retrieval, is 80-years old [4]. The recent decades have been ripe with work that brought a much better understanding of this rich concept. Yet, we still don't know which cognitive and affective processes are involved in relevance judgments. Empirical work that tackles these questions is scarce. This paper aims to contribute toward better understanding of cognitive processing of text documents at different degrees of relevance. Our approach takes advantage of a direct relationship between eye movement patterns, pupil size and cognitive processes, such as mental effort and attention. We examine gaze-based metrics in relation to individual word processing and reading text documents in the context of a constricted information search tasks. The findings indicate that text document processing depends on document relevance and on the user-perceived relevance. Statistical analyses show that relevant documents tended to be continuously read, while irrelevant documents tended to be scanned. Most eye-tracking-based measures indicate cognitive effort to be highest for partially relevant documents and lowest for irrelevant documents. However, pupil dilation indicates cognitive effort to be higher for relevant than partially relevant documents. Classification of selected eye-tracking measures show that an accuracy of 70-75% can be achieved for predicting binary relevance. These results show a promise for implicit relevance feedback.

Categories and Subject Descriptors
H.1.2 User/Machine Systems: Human information processing; H3.3 Information Search and Retrieval: Search process.

General Terms
Measurement, Human Factors.

Keywords
Information relevance, reading, eye-tracking, pupilometry.

1. INTRODUCTION
The main goal of users engaged in information search and retrieval (ISR) is to satisfy their information needs by finding relevant information. The origins of the notion of information relevance go back 80-years [4]. The most recent two-three decades have been ripe with work that brought us better and much richer understanding of this concept. Yet, scholars still lament how little is known about the factors affecting relevance judgments [49] and the discussion on the nature of relevance and its underlying factors continue [21]. In spite of understanding an importance subjective factors play in relevance judgments (e.g., [3, 11]), little work explicitly considers internal psychological factors that are at play in relevance judgments. When such factors are considered, the work tends to be theoretical and lack empirical grounding (e.g., [45]). Consequently, we still don't know "how relevance happens" in human mind and which cognitive and affective processes are involved. Empirical work that tackles these questions has started only recently (e.g., [6, 19, 29, 30, 31]).

Relevance judgments can indicate user's interest and progress in a information search task, and reflect the search system's effectiveness. In IR studies relevance is often measured in terms of explicit user actions, such as display time or saving a document, or through self-assessments. User actions, however, provide ambiguous evidence [24], while subjective assessments can be biased. Ability to infer relevance judgments in a non-intrusive way would provide an objective means to capture this important aspect of the user's mental state while in the 'flow' of search and so enable study of search behavior in natural settings.

We situate our work among these efforts that empirically investigate cognitive processes involved in human-information interaction. Our approach takes advantage of a direct relationship between eye movement patterns and cognitive processes [13]. Spatial–temporal patterns of eye movement linked to document regions and transitions between the documents reflect cognitive processing of the document content and are driven by, among others, document processing in the context of user's information task.

In this work we ask three research questions:

RQ1. Does the degree of relevance of a text document affect how it is read?

RQ2. Does it affect cognitive effort invested in reading it?

RQ3. Could the degree of relevance be plausibly inferred in an non-intrusive way from eye-tracking data?

2. RELATED WORK
We focus the presentation of related work on research projects that employed eye-tracking in relation to information relevance.

2.1 Eye-movement and information search
Eye tracking has received considerable attention as a data source useful in information search and retrieval research. Much of the work using eye tracking data has concentrated on eye fixation patterns on ranked search results pages (e.g., [5, 22, 17, 33]). For example, Guan & Cutrell [18] examined task type influences on user search behavior by manipulating the positions of target results in navigational and informational tasks. Participants took longer to complete tasks and were less successful in finding

IIiX '14, August 26 - 29 2014, Regensburg, Germany

ACM 978-1-4503-2976-7/14/08$15.00.
http://dx.doi.org/10.1145/2637002.2637011

results when the search targets were displayed at lower positions on the results list. The eye tracking data explained that poor performance by showing a decreased probability of looking at search results in the low position.

A number of researchers employed eye-tracking in IR contexts with a focus on inferring relevance and on incorporating this information to improve retrieval results (e.g., [7, 8]). Ajanki et al. [1] used eye-movement based features as implicit relevance feedback. Eye-movement features that significantly contributed to their model included: regressions from following words and relative duration of the first fixation on a word. They showed a modest improvement in mean average precision when the eye-based features were used to select additional query terms. Balatsoukas & Ruthven [2] describe an extensive study of relevance and eye-tracking measures. They showed that users expend more cognitive effort (more frequent and longer fixations) on non-relevant document surrogates. One limitation of their study was the use of only a limited set of eye movement based features (fixation duration and number of fixations), a cut-off of fixations shorter than 200ms, and the use of the talk-aloud method that likely affected the fixation durations [15]. Loboda et al. [28] examined relationship between eye movements and word-level relevance. Only one search topic was used. They found that relevant sentence-terminal words received significantly more fixations than non-relevant words. Buscher at el. [6] reported on two studies in which the relationship between several eye movement measures and text passage relevance was examined. The first reported study examined one topic, the second study examined two topics. They found that the most expressive measure with respect to relevance was based on the length of coherently read text. Surprisingly, they found fixation duration not to be a good discriminator between relevant and not relevant text. Their results showed that eye-gaze based techniques improve performance of an information retrieval ranking algorithm by about 8% overall and by about 27% for poorly performing queries.

Our work is most closely related to [6]. It differs in our use of text documents of varying degrees of relevance, in the use of a larger number of search topics, and in a controlled presentation of documents associated with each task. Furthermore, we use only fixations that are part of reading sequences and eliminate single lexical fixations.

2.2 Pupilometry

Pupil dilation is controlled by the Autonomic Nervous System (ANS) [32]. Under constant illumination it has been associated with a number of cognitive functions, including mental effort [23], interest [25], surprise [36], and making a decision. All the sources of variation in pupil's size are related to attention [20, 44].

It is reasonable to expect that degree of document relevance will affect level of attention and, thus, pupil diameter. Yet, the only published work that examined pupil size in relation to information relevance is research by Oliveira et al. [31]. For text documents and images, the authors reported that pupil dilated for higher relevance stimuli.

3. METHOD

We conducted an experiment, in which participants (N=24; 9 females) were asked to find information in short text documents containing news stories. Participants were recruited from undergraduate and graduate student body at Rutgers University; 66.7% were 18-22 years old, 16.7% 23-27, and the remaining 16.7% older than 32. All participants reported English as their first language and were screened for normal to corrected-to-normal vision. Each participant received $25 compensation. The experiment was approved by the university's IRB.

3.1 Experimental setup.

The experiment was conducted in a lab equipped with a PC computer running Windows XP and connected to Tobii T60 remote eye-tracker. The T60 is integrated into a 17-inch TFT monitor with a resolution of 1280x1024 pixels (Tobii Technology, Stockholm, Sweden). Participants were seated approximately 65-75cm from the monitor. The text documents were displayed using 19 points Verdana font. A line of text was 32 pixels high (between the ascender and descender lines [50]), which corresponds to about 0.9^{o} of visual angle. The text was displayed within a rectangle of 1020x900 pixels. This was done to avoid using at least 100 pixels at the left, right, and bottom edge of screen, as the eye-tracker's accuracy is typically lower close to the screen edges.

3.2 User tasks and procedure

The experiment had within-subject design. Each participant performed two types of tasks: 1) target word search (WS task), and 2) a constricted information search (IS task). Each session included 21 trials of each task type presented in randomized order. We decided on 21 trials to include as wide a range of information topics possible, while being constrained by the pragmatic limitation of session length. Before the actual task began, participants performed two training trials. We focus this paper on reporting results from the IS task. The essential elements of experimental design are shown in Figure 1. The IS task involved finding factual information in news stories. Information finding was prompted by a question informing participants what information they were expected to find in the subsequently presented documents. An example question: "Which Russian fleet was submarine Kursk part of?" We call this task a constricted information search, as participants did not enter queries but, instead, after seeing a question they were presented with documents as if a search had already been performed and they were viewing individual search results.

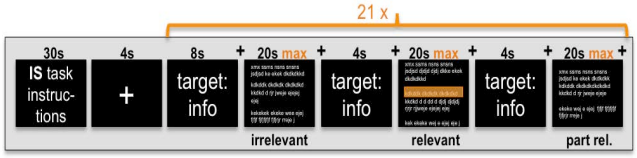

Figure 1. Experimental Design

During an experiment session, first, general task instructions were presented on screen for 30 seconds, next a fixation screen appeared for 4 seconds, then a question was displayed for 8 seconds. To remind participants of the current question, it was repeated briefly (4s) before the second and third document (shown in Figure 1 as "target info"). Fixation screens were presented for 4 seconds before each question or text stimulus (after the first fixation screen occurrence in Figure 1, fixation screens are represented as "+" above the stimuli). Participants were asked to judge document relevance on a binary scale (yes/no). Binary scale was appropriate since the participants were instructed to find documents which contained answer to the question. In fourteen "overt" trials they responded explicitly by pressing one of two buttons marked "yes" or "no" on a keyboard. These explicit ratings of document relevance by participants are referred to as perceived relevance. In the seven remaining "covert" trials

participants were instructed to make their relevance decision "in their heads" only, without expressly pressing "yes"/"no" buttons; once they made their decision they moved on to the next document by pressing the space key.

The selected twenty one questions were presented in randomized order; each question was followed by three news stories of varied relevance: Irrelevant (I) – a document on a topic different from the question, Topical, or partially relevant (T) – a document on the question's topic, but not containing an answer to the question, and Relevant (R) – a document containing an answer to the question. Thus, each participant saw sixty-three news stories. The order of document relevance levels within each trial was pseudo-randomized using the following process. A set of three documents in each trial was created to contain exactly one relevant document and a combination of topical or irrelevant documents. Thus, within each trial there were three possible combinations of relevance levels RTT, RTI, RII. These combinations can be permuted in 3, 6, and 3 ways respectively, yielding the following twelve orders: RTT, TRT, TTR, RTI, RIT, IRT, ITR, TRI, TIR, RII, IRI, IIR. These twelve orders were used to create a sequence of 21 trails (9 of them were repeated twice). We followed this procedure in order to avoid exposing participants to similar orders of document relevance, from which they could plausibly learn patterns and thus start guessing relevance levels.

3.3 Document set

Text documents used in the study came from a large set of news stories available as the AQUAINT corpus [16]. All news stories were in English and originated from several international sources, such as Associated Press, New York Times, and Xinhua. We selected a small subset of stories (initially about 200) aiming to achieve a relatively low variation in the text length. We obtained task questions along with their relevance assessments for the documents from TREC 2005 question answering (Q&A) task [41]. We further selected 21 documents for WS task, 63 for IS task and 2 for trainings tasks and manually verified the relevance assessments for the selected document subset. Given the simplicity of the selected question answering fact-finding tasks, we did not need to perform inter-rater agreement assessment. These ratings constitute relevance ground truth and are referred to as *document relevance*.

We processed the news stories into experimental stimuli using the following procedure. First, we created HTML documents from the new stories to automate capture of AOIs for each word as it was rendered on screen. Second, we marked up relevant words (i.e. words that contained an answer to the question) in relevant documents and stored these words along with their co-ordinates in a database. Third, we took a screen shot of each HTML document (without any markup of the relevant words) and used the resulting images as text stimuli (screen resolution was maintained the same).

3.4 Eye-tracking measures

Eye movements were analyzed using our reading modeling approach described in [10], which is briefly summarized below. Our approach is influenced by the E-Z Reader model [37, 38]. Assumptions of that model that are pertinent to our work: 1) reading is serial and words are processed one at a time in the order of their appearance in text, 2) more than one word can be processed on a fixation, because next word can be processed in parafoveal view, and 3) there is a minimum fixation time required for acquisition of a word's meaning. Accordingly, we use fixation duration threshold of 150ms. We refer to fixations above this threshold as lexical fixations. We implemented a simple, two-state, line-oriented reading model in Python and used it to group lexical fixations into reading and scanning sequences (Figure 2). A scanning state represents isolated lexical fixations. A reading state represents reading in one line. If reading continues to the subsequent line, it is currently represented by a new reading state.

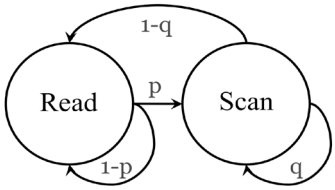

Figure 2. Two-state modeling of reading.
p and q are probabilities of transitions between states.

To investigate our second research question, we used reading-fixation-sequence based measures that have been suggested as reflecting aspects of cognitive effort [35, 38]:

- fixation duration,
- number of regression fixations in the reading sequence,
- the spacing of fixations in the reading sequence (called *perceptual span*),
- reading speed, a) defined as the length of reading in pixels per unit time, b) as the number of words fixated upon per unit of time, and
- reading length, a) defined as the length of reading in pixels; b) defined as the number of words fixated upon.

We also included reaction time (RT) as a measure of cognitive effort (often used in psychology and human factors [43]). Longer reaction times indicate more effort involved in accomplishing a task. RT was defined in our study as the time from the onset of document presentation to the participant's key press expressing their relevance judgment. Reaction time is equivalent to dwell time on a document. Dwell time was used in information search and retrieval as an indicator of user's interest [27, 26], but it was also criticized as a measure insufficient on its own [24].

3.5 Pupil measures

As mentioned in section 2.2, changes in pupil diameter are indicative of a variety of cognitive processes, but also of changing lighting conditions in the surrounding environment [32, 48]. Controlled lab environment with constant lighting, and the use of text documents with black background and a similar number of words presented in white font ensures virtually constant luminance across all experiment sessions. To eliminate individual variability in pupil sizes, we obtained a baseline for each participant by taking an average pupil size over all text document stimulus presentations. We could have obtained pupil baselines from fixation screen stimuli that preceded each text display (similarly as it was done by [31, 48]), however, the fixation screens were darker than the text stimuli since the former contained only a white cross, while the latter contained an average of 178 words displayed in a white font. Similarly as in [48], the relative change in pupil diameter (pr^i_t) was calculated as a difference between pupil measurement at a time t (p_t) and the average size of pupil for the participant i (p^i_{avg}) normalized by that average, that is:

$$pr^i_t = (p_t - p^i_{avg})/p^i_{avg} \qquad (1)$$

4. RESULTS

4.1 Data cleanup and pre-processing

Data analysis began with cleaning up the eye-tracking data. We used only those fixations where the data quality was good and where fixation was within the screen coordinates. Bad quality fixations were defined as missing eye (reported by Tobii eye-tracker as validity=4). This resulted in removing approximately 5% of fixations. We then considered continuity of eye-tracking data and, following procedure described in [31], removed trials with gaps longer than 1 second. We also removed trials with over 20% of eye-tracking data missing and with less than 4 fixations on a document. This process left small gaps in eye-tracking data (perhaps due to eye blinks), to eliminate them we "merged" pupil measurements for both eyes by calculating their average, if both were captured, or by using pupil measurement of one of the eyes, if the other eye was "lost". We then applied MatLab's (Mathworks, Natick, MA) linear interpolation algorithm to the remaining gaps in pupil data.

The data cleanup steps resulted in removing approximately 10% of trials. The number of trials that were left was not equally distributed between the three relevance levels, I/T/R. To obtain a more uniform distribution of document relevance degrees and their text lengths, we removed data from irrelevant documents longer or equal 310 words (the word limit was established from our document characteristics). This yielded average length of documents at 178 words (SD=30). After this operation was completed, we obtained the following distribution of document relevance levels, I/T/R: 19/21/19, and for combinations of document relevance/trials instances, I/T/R: 259/206/265. The difference of about 20% is acceptable, given the robustness of the analysis of variance.

We removed the typical individual variability of eye-tracking measures by performing two procedures. Section 3.5 describes how we did it for pupil diameter measurements. For all other eye-tracking measures we calculated z-scores. The underlying procedure is similar to measure personalization described in [6].

Most of the trials that were removed due to bad eye-tracking data happened to be the "covert" trials. Consequently, we present data analysis for "overt" trials, in which participants responded explicitly by pressing yes/no. However, we also performed analysis that included both types of trials and we generally obtained similar relationships, only the statistical significance and effect size tended to be smaller for "covert" trials.

4.2 Independent and dependent measures

Dependent measures were obtained directly or derived from data captured by the eye-tracker (described in sections 3.4, 3.5). Independent measures were *document relevance* and *perceived relevance*.

We performed two types of data analysis, statistical analyses using SPSS v.18 and classification using Weka 3.7 [46]. The statistical analyses involved a series of one-way ANOVAs, either with document relevance or with perceived relevance as an independent factor. These analyses examined the effects of relevance on reading vs. scanning, the number and duration of reading sequences, fixation duration, regressions, perceptual span, reading length and speed and on pupil dilation. Classifications were then performed for combinations of these variables.

4.3 Participant relevance judgment accuracy

We first report cross tabulation of document relevance and participants' perceived relevance.

Table 1. Document relevance vs. perceived relevance

		Document Relevance			
		I	**T**	**R**	**Total**
Perceived Relevance	**I**	**258**	**178**	36	470
	R	2	29	**229**	260
	Total	260	205	265	730
	Mean participant accuracy	99.2%	85.9%	86.4%	Overall accuracy: 90.3%

Contents of cells with correct answers are bolded.

As expected, for our simple task, the participant accuracy of relevance judgments was high.

4.4 Reading model & eye movement patterns

4.4.1 Statistical analysis

The two-state reading model is shown in Figure 2. In the following analyses we examine differences in transition probabilities between the two states of our model and document relevance and perceived relevance. Table 2 presents results from two separate one-way ANOVAs, one for document relevance, and the other for perceived relevance.

Table 2. Probabilities of transition between reading and scanning states. (S-scanning, R-reading)

	Transition	**I**	**T**	**R**	**ANOVA**
Document relevance	SS	0.52(.013)	0.46(.013)	0.37(.017)	F(2,470)=27
	SR	0.48(.016)	0.54(.016)	0.63(.016)	F(2,470)=27
	RR	0.78(.007)	0.84(.007)	0.88(.006)	F(2,461)=90
	RS	0.22(.007)	0.16(.007)	0.12(.006)	F(2,461)=90
Perceived relevance	SS	0.49(.011)	n/a	0.36(.016)	F(1,457)=53
	SR	0.51(.011)	n/a	0.64(.016)	F(1,457)=53
	RR	0.81(.005)	n/a	0.88(.006)	F(1,648)=166
	RS	0.19(.005)	n/a	0.12(.006)	F(1,648)=166

Mean (std. error). Due to lack of homogeneity of variance, we report in this table Welch's corrected F. In all cases p<.001. Post hoc tests (Games-Howell, in which equality of variance is not assumed), performed for the three *document relevance* levels, showed significance for all pairwise comparisons of each variable.

The highest probability of reading was found for relevant documents (Table 2), while the highest probability of scanning was found for irrelevant documents. Topical documents were in the middle. The same pattern was observed for perceived relevance. An illustrative example of differences in eye movement patterns between for three selected documents at different levels of relevance is shown in **Figure 3**.

4.4.2 Classification results

We used decision tress C4.5 algorithm (as implemented in Weka 3.7 under the name J48 [46]) with 10-fold cross-validation to examine if document relevance and, separately, perceived relevance, can be predicted by the probabilities of transitions between the reading states. For all classifications with three classes, the chance accuracy is 33%, while for two classes, it is 50%. An application of a classification algorithm to our data that yields accuracy levels above these baseline values indicates an improvement.

Table 3. Classification of relevance using transition probabilities.

Kind of Relevance and Its Classes		Accuracy
Document relevance	3-classes: R,T,I	52%
	2-classes: R,I; T merged with I	67%
	2-classes: R,I; T merged with R	70%
Perceived relevance - 2-classes: R,I		72%

4.5 Eye-fixation-derived cognitive effort measures

4.5.1 Statistical analysis

We investigated if variables described in section 3.4 differed between levels of document relevance and perceived relevance. **Table 4** and **Table 5** each presents results from two separate one-way ANOVAs, one for document relevance, and the other for perceived relevance.

Table 4. Eye-tracking derived cognitive effort measures. Calculated per document and normalized by document length in words.

	Document relevance			*Perceived relevance*	
Variable	**I**	**T**	**R**	**I**	**R**
reaction time (RT) [ms/word]	42** (1.4)	71** (1.7)	59** (1.6)	55 (1.3)	56 (1.6)
length of all reading sequences [px/word]	8.5** (.4)	16** (.6)	14** (.5)	12* (.4)	13.5* (.5)
duration reading [ms/word]	25** (1.1)	47** (1.6)	41** (1.33)	35** (1.1)	40** (1.3)
duration scanning [ms/word]	12** (.43)	15** (.64)	9** (.42)	13** (.4)	8.7** (.4)
number of reading sequences [per word]	.05** (.002)	.08** (.002)	.07** (.002)	.06 (.002)	.06 (.002)
total number of fixations [per document word]	.16** (.005)	.27** (.007)	.22** (.006)	.21 (.005)	.27 (.007)

Mean (std. error). *significant at p<.01; **significant at p<.001. Post hoc tests (Games-Howell) showed that all pairwise comparisons for each significant variable between *document relevance* levels were significant at p<.001.

Normalized reaction time (i.e. time to from the onset of text stimulus presentation to the participant's key press expressing its relevance judgment; normalized to the document's length in words) can be considered a simple measure of cognitive effort. This variable shows that judging topical documents was more effortful, while judging irrelevant documents was the easiest (**Table 4**). More interestingly, eye-tracking-based measures generally show a similar pattern. This applies to measures normalized by document's length in words (**Table 4**) as well as to most measures calculated at the document level (**Table 5**). Exceptions to this pattern is reading speed in pixels (**Table 5**) and duration of the longest reading sequence as well as the maximum fixation duration. The reading speed in pixels variable indicates that reading irrelevant documents was slower, and reading topical or relevant documents was faster. We note that there was no difference in reading speed expressed in words between the different documents. The duration variables are the highest for the relevant documents (**Table 5**). The patterns are generally similar for document relevance as for perceived relevance.

Table 5. Eye-tracking derived cognitive effort measures Calculated per document.

	Document relevance			*Perceived relevance*	
Variable	**I**	**T**	**R**	**I**	**R**
reaction time (RT) [s]	7.3** (.2)	12.3** (.3)	9.8** (.25)	9.5 (.21)	9.8 (.26)
reading speed in pixels [px/s]	190** (5)	223** (6)	239** (5)	203** (4)	233** (5)
reading speed in words [words/s]	3.3 (.04)	3.4 (.04)	3.3 (.04)	3.4 (.03)	3.4 (.04)
duration of the longest reading sequence [s]	1.4** (.05)	1.9** (.07)	2.1** (.7)	1.6** (.04)	2.2** (.07)
max fixation duration in a reading sequence [ms]	415** (17)	448** (8)	494** (10)	430** (10)	496** (10)
number of words fixated upon	21** (.6)	35** (.8)	27** (.8)	28 (.6)	27 (.8)
number of fixations on words	24** (.8)	42** (1.1)	34** (1.0)	32 (.8)	33 (1)
proportion of words fixated on [%]	13%** (.04%)	20%** (.05%)	16%** (.05%)	17% (.05%)	16% (.07%)

Mean (std. error). **significant at p<.001. Post hoc tests (Games-Howell) showed that pair-wise comparisons of each significant variable between *document relevance* levels were significant at p<.001, with the following exceptions: a) the difference between duration of the longest reading sequence was significant at p=.029 for T and R documents; b) no-significant difference between the reading speed in pixels for T and R documents.

To help reader see patterns in **Table 4** and **Table 5**, the following table (**Table 6**) summarizes the results from those two tables.

Table 6. Patterns of reading and scanning activities derived from cognitive effort measures.

	Document relevance			*Perceived relevance*	
Construct	**I**	**T**	**R**	**I**	**R**
cognitive effort reading activity	L	H	M	L	H
scanning activity	M	H	L	H	L

H – high, M – medium. and L – low values.

The ratio of words fixated upon (that is unique words) and all fixations on words reflect proportion of words that were fixated on again (that is, the number of regressions). For irrelevant documents, that proportion is 10%, for topical documents 20%, and for relevant documents it reaches 25%. Comparing these values with the typical value of 10%-15% for regressions in reading [37], we note that reading topical, and, in particular relevant documents involved re-reading a higher than expected proportion of words. Examining selected sequences of eye movements (scan paths) on relevant documents (see for example **Figure 3**), we found that the relevant words or phrase was typically read again just before a relevance judgment was made.

The reported analyses were performed at the aggregate level of all participants. To check if the eye-tracking data from individual participants followed similar patterns, we performed separate analyses at the level of individual participants. The individual analysis showed in most cases (for 80-96% of participants) and for most normalized variables, patterns similar to the results of analysis across all participants presented in **Table 4** and **Table 5**,. For a duration of scanning the agreement was somewhat lower

(for 71% of participants). In all cases, the number of individual participants for whom the patterns were confirmed, was greater than the chance probability (50%).

4.5.2 Classification results

We used the same algorithm as in previous classification. For the cognitive effort eye-tracking measures, however, we first applied a supervised attribute selection (BestFirst algorithm in Weka).

Table 7. Classification of relevance using cognitive effort eye-tracking measures from Table 4 **and** Table 5 **(except RT).**

Kind of Relevance and Its Classes		Accuracy
Document relevance	3-classes: R,T,I	54%
	2-classes: R,I; T merged with I	71%
	2-classes: R,I; T merged with R	74%
Perceived relevance - 2-classes: R,I		72%

We explored several other classification algorithms to find a better performing one. The best classification accuracy (75%) for perceived relevance was obtained by support vectors for binary classification (SMO) [34]. It was an improvement of 3% over the decision trees algorithm, we did not find such improvements in classification for any other classification we report in this paper.

Table 8. Classification of relevance using reaction time.

Kind of Relevance and Its Classes		Accuracy
Document relevance	3-classes: R,T,I	52%
	2-classes: R,I; T merged with I	64%
	2-classes: R,I; T merged with R	72%
Perceived relevance - 2-classes: R,I		64%

Classification using reaction time normalized to document length in words yielded the same accuracies, expect for document relevance 3-class classification that was only 49% accurate.

Table 9. Classification of relevance using all variables (except for pupil dilation)

Kind of Relevance and Its Classes		Accuracy
Document relevance	3-classes: R,T,I	55%
	2-classes: R,I; T merged with I	72%
	2-classes: R,I; T merged with R	74%
Perceived relevance - 2-classes: R,I		74%

4.6 Pupil measures

The Tobii T-60 eye-tracker captures eyes at a frequency of 60Hz. That is, every 16.67ms a new measurement of pupil diameter is performed. To avoid information loss inherent in calculating mean values, we decided to treat all pupil measurements as individual values. The relative pupil dilation (pr) was calculated using equation (1).

4.6.1 Statistical analysis

We observed the largest pupil dilation for relevant documents and pupil contraction for irrelevant documents (Table 10). These effects are particularly apparent for changes in pupil size in the last second before a participant's relevance judgment. The patterns of changes in pupil diameter are different from those of other eye-tracking measures that reflect cognitive effort in reading. The pupil size is still the lowest for irrelevant documents, but it is the highest for the relevant documents and not for the partially relevant ones.

However, the observed effects of relevance on changes in pupil size were small. Thus, in future work a more sophisticated analysis should be applied that takes into account how fast pupil dilates, and uses, for example, frequency domain or principal component analysis (PCA) of pupillary changes [31, 32]

Table 10. Relative pupil dilation measured during the whole text stimulus exposure

	I	T	R	ANOVA
Document relevance	- 1.2% (.02%)	- 0.4% (.02%)	+0.8% (.02%)	F(2, 505183)=3439
Perceived relevance	- 0.9% (.02%)	n/a	+1.1% (.02%)	F(1,505186)=8689
Relative pupil dilation measured during the last 1000ms before a participant's response				
Document relevance	- 0.3% (.05%)	+1% (.05%)	+2.9% (.05%)	F(2, 47498)=1211
Perceived relevance	+0.1% (.03%)	n/a	+3% (.05%)	F(1, 47499)=2696

Mean (std. error). Post hoc tests (Games-Howell) showed that all pair-wise comparisons of each significant variable between *document relevance* levels were significant at p<.001.

4.6.2 Classification results

The classification was performed on over 500,000 cases of pupil measurements labeled with document relevance or perceived relevance.

Table 11. Classification of relevance using relative pupil dilation measured during the whole text stimulus exposure

Kind of Relevance and Its Classes		Accuracy
Document relevance	3-classes: R,T,I	37%
	2-classes: R,I; T merged with I	65%
	2-classes: R,I; T merged with R	70%
Perceived relevance - 2-classes: R,I		65%
Table 9. (cont.) Relative pupil dilation (pr) measured during the last 1000ms before participant's response		
Kind of Relevance and Its Classes		**Accuracy**
Document relevance	3-classes: R,T,I	46.5%
	2-classes: R,I; T merged with I	67%
	2-classes: R,I; T merged with R	63%
Perceived relevance - 2-classes: R,I		67%

5. DISCUSSION

We structure the discussion by revisiting our three research questions.

RQ1. Does the degree of relevance of a text document affect how it is read?

The results indicate that the degree of relevance of a text document does affect how it is read. Significant differences in reading patterns were found between documents at the three levels of relevance. The results indicate that relevant documents tended

to be continuously read, while irrelevant documents tended to be scanned. Example patterns are shown in **Figure 3**. Our findings generally agree with [6] in that relevant documents tend to be read more coherently, whereas irrelevant documents tend to be scanned.

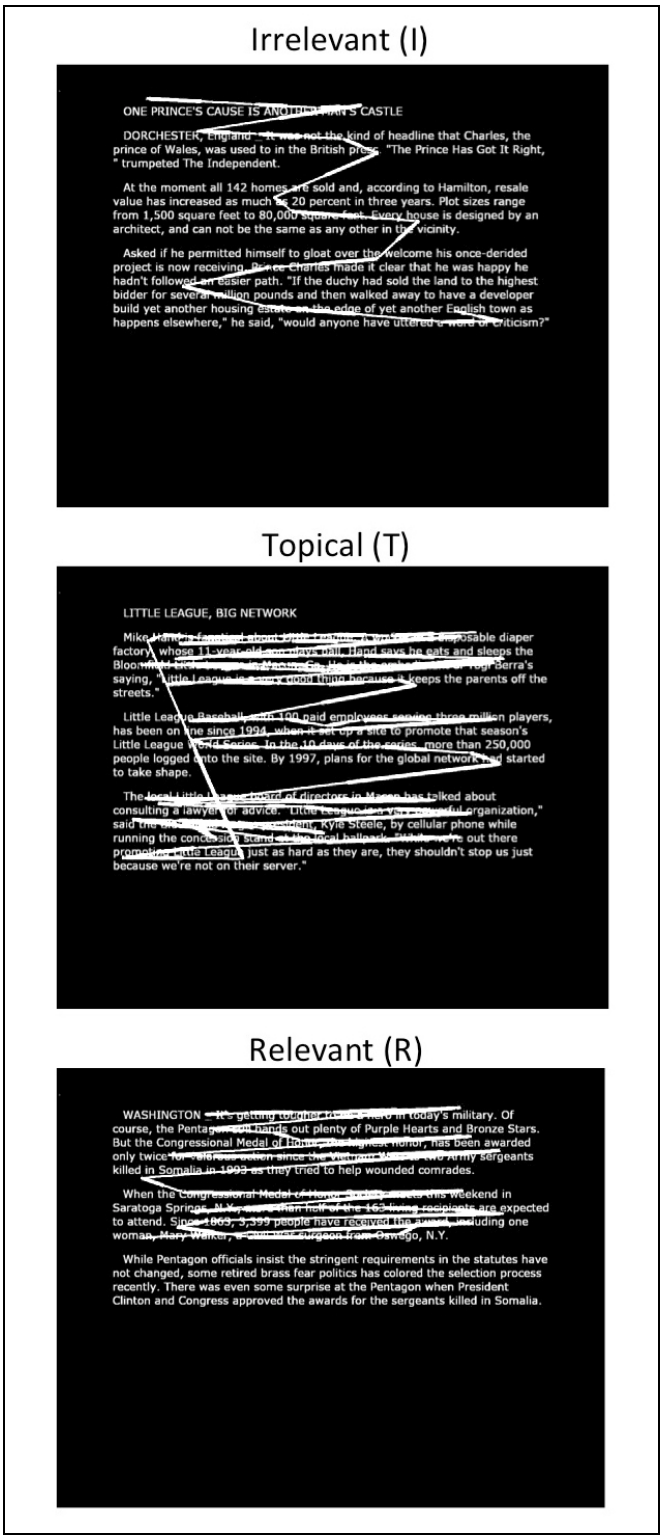

Figure 3. Example eye movement patterns for documents at different levels of relevance (experiment stimuli screenshots).

Our results also show a difference in re-reading words between the documents at different level of relevance. For relevant documents, about 10-15% more words were re-visited beyond what is implied by reading models [37]. For partially relevant documents the increase in re-reading was 5-10%. Re-reading is one aspect of increased cognitive demands imposed by processing partially or fully relevant documents. This finding leads to the discussion of our next research question.

RQ2. Does it affect cognitive effort invested in reading it?

The degree of relevance of a text document seems to affect cognitive effort involved in reading it. In most cases, cognitive effort inferred from eye-tracking data was highest for partially relevant documents and lowest for irrelevant documents. Reaction time and a collection of eye-tracking based measures indicate that the lowest cognitive effort was involved in judging that a news story is not relevant. Judging topically relevant documents required highest effort, while the effort involved judging relevant documents was generally in the middle. These results agree with Villa and Halvey [40], who used subjective workload judgment (NASA TLX) to investigate effort involved in relevance judgment. Their results show the same direction of relationship between effort and judging irrelevant and topical documents. However, they did not find a significant differences in cognitive effort between judging irrelevant and relevant documents. This difference could be because we employed a different and a more sensitive assessment of cognitive effort.

Examining the absolute measures, the duration of the longest reading sequences and the longest fixation in reading sequences (**Table 5**), we found that they were longest for the relevant documents. This is likely an indication that maximum peak of cognitive demands [47] occurred while reading and processing some parts of a relevant document, but that, on the average, the effort involved in processing these documents was lower than involved in processing topical documents.

The observed changes in pupil dilation followed the pattern of absolute measures. This may be an indication that changes in pupil size obtained from its discrete measurements reflect similar aspects of cognitive effort as the absolute measures do. It may simply be a result of some more aggregation at the trial level in the processing of other eye-tracking measures. It may also mean that different measures reflect somewhat different aspects of cognitive effort.

RQ3. Could the degree of relevance be plausibly inferred in an non-intrusive way from eye-tracking data?

Our classification results (shown in Table 3, Table 7, Table 8, Table 9, and Table 11) suggest a feasibility of using eye-tracking based measures to predict text document relevance, and thus a user's intent. Prediction accuracy levels obtained for the three-class document relevance ranged from 37%-54%. The obtained accuracy levels are above the chance levels. For two-class predictions, the prediction accuracies ranged from 65% to 75%.

The best group of predictor variables were the eye-tracking derived cognitive effort measures reaching prediction accuracy of 75% for binary relevance (Table 7). Using all eye-tracking variables (expect for pupil dilation) just slightly improved the classification across all categories by 1-2% (Table 9).

Although, implicit relevance feedback (IRF) has not been in our current focus, we can compare our results to work of Moshfeghi

and Jose [29]. The task used in our experiment is comparable to their INS task that simulated the information seeking search intent. They achieved the best accuracy of 67.66% (cf. [29] Table 2) by using a combination of reaction time (dwell time) and facial expressions. Our eye-tracking based measures of cognitive effort seem to offer a somewhat higher accuracy (up to 75%) for inferring binary document relevance, while our other measures yielded accuracy at levels similar to theirs. We cannot compare our results to other search intentions used in their study, such as the entertainment-based search intent, where their achieved accuracy was up to 82.66% (using dwell time and heart rate).

It is interesting to note patterns of classification accuracy for cases where document relevance is forced to two classes by changing class T to I (TI), or, separately, class T to R (TR). The first case (TI) corresponds to the expected rating of partially relevant documents as irrelevant, since these documents did not contain answers to the questions. Since most partially relevant documents were judged by participants as irrelevant (which corresponded to our experimental design), document relevance case TI is similar to perceived relevance. This is corroborated by similar levels of accuracy achieved for both of them in case of cognitive effort eye-tracking measures, reaction time, and relative pupil dilation. The second case (TR) represents classification of documents based on their semantic similarity and, thus, without regard to the task used in this experiment. One could draw a rather speculative conclusion that the variables for which higher accuracy was obtained for case TR may be indicators of similarities in cognitive processing of partially relevant documents to relevant documents. The good news is that these measures may work well for tasks where topical relevance is an important criterion. The bad news is that these measures are task dependent and work less well for fact finding tasks (as used in this experiment) and for predicting perceived relevance on such tasks.

6. CONCLUSIONS AND FUTURE WORK

Overall, our results demonstrate that the degree of relevance of a text document does affect how it is read and that it does affect the level of cognitive effort required to read documents.

Our results largely agree with prior findings. However, our contribution is not just in confirming prior results, but also in extending them to documents with three levels of relevance and to a wider range of information topics. The latter is a likely indication that the relationships are independent of topics. This conclusion is corroborated by results from ANOVA with question topic as a factor. While the topic effect was overall significant, the post-hoc tests showed that this effect was, in most cases, due to just one topic.

One limitation of the study lies in our construction of reading states. Continuous reading in each line is considered a separate reading state. This limitation, however, does not impact the measures used in this paper. Another limitation is the use of binary ratings of perceived relevance, while using three-level document relevance ratings. However, forcing participants to make binary judgments corresponded well with IS task used in the experiment. The simple and constricted task employed in this experiment limits the extent to which we can generalize our results. Another limitation is the relative homogeneity of the text documents used in our study. The documents were of the same genre (news stories) and were selected to have similar length. Thus, they were at a similar readability level [39]. From prior research [9], we know that text readability level may affect relevance judgments, and, hence, that it should be considered as one of the factors in studies of cognitive aspects of relevance assessment. Our intention was to start with a simple and well understood task, and with a well controlled document genre. We plan to extend task types and document genre in future work.

The current eye-tracking technology has its own limitations that impact data collected by these instruments. The Tobii T-60 eye-tracker has factory reported accuracy of 0.5^{o} visual angle. It impacts the ability to detect individual words on which eyes fixate. As described in the Method section, we took several measures to ensure accuracy of collecting eye fixations on words.

In this paper, we investigated the differences in reading patterns and in cognitive effort between documents of different degrees of relevance. We also attempted to use our data to classify document relevance. One interesting question for future research is to examine how much of eye-tracking data on a given document is needed to reliably predict the document's degree of relevance. This project focused on eye movements on short text documents, in the follow-up work, we plan to examine other information objects, for example, search engine result pages (SERPs) and search results overviews (e.g., lists, tag clouds). We are also working on combining psycho-physiological signals from other sources (e.g., EEG) and plan to use them together with eye movement data. The data will be combined at the level of individual words (where we plan to use eye-fixation-related potentials (EFRPs) and extend recent work by Frey et al. [14] and by Eugster et al. [12]), as well as at the level of search result surrogates, document sections (e.g., paragraphs) and whole documents. Future work will also examine a wider range of properties of pupil dilation signal and additional methods of its analysis (e.g., speed of change, principal components analysis, and frequency analysis). We will look at changes of cognitive effort while a user is reading a document, before and after she encounters the relevant words.

Our goals in this research are to contribute to better understanding of differences in cognitive processing of text documents at different levels of relevance in relation to a user's information needs. By doing so we eventually aim to make an applied contribution to the methods of implicit relevance feedback [42] and to information retrieval systems that can infer user's search intent. An information retrieval systems that knows perceived document relevance can, for example, use this information and return a document set that closer matches a user's information need. We also intend to make a contribution at the theoretical level. In spite of the wealth of research on relevance judgments, little is known on internal psychological factors that are at play when searchers are making relevance assessments. Through our use of neurophysiological data combined with interaction logs and searchers' verbal explanations, we aim to investigate these factors in order to improve our understanding of cognitive and affective processes involved in relevance judgments and of "how relevance happens" in human mind.

7. ACKNOWLEDGMENTS

This research project was funded, in part, by the Google Faculty Research Award and by the IMLS Career Development Grant #RE-04-11-0062-11A. We thank Michael Cole for contribution to the experimental design and for his help with conducting experimental sessions. We are also grateful to the anonymous reviewers for their helpful comments that allowed us to improve the paper.

8. REFERENCES

1. Ajanki, A., Hardoon, D., Kaski, S., Puolamäki, K., and Shawe-Taylor, J. Can eyes reveal interest? Implicit queries from gaze patterns. *User Modeling and User-Adapted Interaction 19*, 4 (2009), 307–339.
2. Balatsoukas, P. and Ruthven, I. An eye-tracking approach to the analysis of relevance judgments on the Web: The case of Google search engine. *Journal of the American Society for Information Science and Technology 63*, 9 (2012), 1728–1746.
3. Borlund, P. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology 54*, 10 (2003), 913–925.
4. Bradford, S.C. Sources of information on specific subjects. *Engineering: An Illustrated Weekly Journal (London) 137*, 26 january (1934), 85–86.
5. Brumby, D.P. and Howes, A. Strategies for Guiding Interactive Search: An Empirical Investigation Into the Consequences of Label Relevance for Assessment and Selection. *Human–Computer Interaction 23*, 1 (2008), 1–46.
6. Buscher, G., Dengel, A., Biedert, R., and Elst, L.V. Attentive documents: Eye tracking as implicit feedback for information retrieval and beyond. *ACM Trans. Interact. Intell. Syst. 1*, 2 (2012), 9:1–9:30.
7. Buscher, G., Dengel, A., and van Elst, L. Eye movements as implicit relevance feedback. *CHI '08 extended abstracts on Human factors in computing systems*, ACM (2008), 2991–2996.
8. Buscher, G., Dengel, A., and van Elst, L. Query expansion using gaze-based feedback on the subdocument level. *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2008), 387–394.
9. Chandar, P., Webber, W., and Carterette, B. Document Features Predicting Assessor Disagreement. *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM (2013), 745–748.
10. Cole, M.J., Gwizdka, J., Liu, C., Bierig, R., Belkin, N.J., and Zhang, X. Task and user effects on reading patterns in information search. *Interacting with Computers 23*, 4 (2011), 346–362.
11. Cosijn, E. and Ingwersen, P. Dimensions of relevance. *Information Processing & Management 36*, 4 (2000), 533–550.
12. Eugster, M.J.A., Ruotsalo, T., Spapé, M.M., et al. Predicting Term-relevance from Brain Signals. *Proceedings of the 37th International ACM SIGIR Conference on Research & Development in Information Retrieval*, ACM (2014), 425–434.
13. Findlay, J.J.M. and Gilchrist, I.D. *Active Vision: The Psychology of Looking and Seeing*. Oxford University Press, Incorporated, 2003.
14. Frey, A., Ionescu, G., Lemaire, B., Lopez-Orozco, F., Baccino, T., and Guerin-Dugue, A. Decision-making in information seeking on texts: an Eye-Fixation-Related Potentials investigation. *Frontiers in Systems Neuroscience 7*, 39 (2013).
15. Gerjets, P., Kammerer, Y., and Werner, B. Measuring spontaneous and instructed evaluation processes during Web search: Integrating concurrent thinking-aloud protocols and eye-tracking data. *Learning and Instruction 21*, 2 (2011), 220–231.
16. Graff, D. The AQUAINT Corpus of English News Text. 2002. http://www.language-archives.org/item/oai:www.ldc.upenn.edu:LDC2002T31.
17. Granka, L.A., Joachims, T., and Gay, G. Eye-tracking analysis of user behavior in WWW search. *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*, ACM (2004), 478–479.
18. Guan, Z. and Cutrell, E. An eye tracking study of the effect of target rank on web search. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2007), 417–420.
19. Gwizdka, J. Looking for Information Relevance In the Brain. *Gmunden Retreat on NeuroIS 2013*, (2013), 14.
20. Hoeks, B. and Levelt, W.J.M. Pupillary dilation as a measure of attention: a quantitative system analysis. *Behavior Research Methods, Instruments, & Computers 25*, 1 (1993), 16–26.
21. Huang, X. and Soergel, D. Relevance: An improved framework for explicating the notion. *Journal of the American Society for Information Science and Technology 64*, 1 (2013), 18–35.
22. Joachims, T., Granka, L., Pan, B., Hembrooke, H., Radlinski, F., and Gay, G. Evaluating the accuracy of implicit feedback from clicks and query reformulations in Web search. *ACM Trans. Inf. Syst. 25*, 2 (2007).
23. Kahneman, D. and Beatty, J. Pupil Diameter and Load on Memory. *Science 154*, 3756 (1966), 1583–1585.
24. Kelly, D. and Belkin, N.J. Display Time As Implicit Feedback: Understanding Task Effects. *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM (2004), 377–384.
25. Krugman, H.E. Some applications of pupil measurement. *JMR, Journal of Marketing Research (pre-1986) 1*, 000004 (1964), 15.
26. Liu, C., Liu, J., Belkin, N., Cole, M., and Gwizdka, J. Using dwell time as an implicit measure of usefulness in different task types. *Proceedings of the American Society for Information Science and Technology 48*, 1 (2011), 1–4.
27. Liu, J., Gwizdka, J., Liu, C., and Belkin, N.J. Predicting task difficulty for different task types. *Proceedings of the American Society for Information Science and Technology 47*, 1 (2010), 1–10.
28. Loboda, T.D., Brusilovsky, P., and Brunstein, J. Inferring word relevance from eye-movements of readers. *Proceedings of the 16th international conference on Intelligent user interfaces*, ACM (2011), 175–184.
29. Moshfeghi, Y. and Jose, J.M. An effective implicit relevance feedback technique using affective, physiological and behavioural features. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM (2013), 133–142.

30. Moshfeghi, Y., Pinto, L.R., Pollick, F.E., and Jose, J.M. Understanding Relevance: An fMRI Study. In P. Serdyukov, P. Braslavski, S.O. Kuznetsov, et al., eds., *Advances in Information Retrieval*. Springer Berlin Heidelberg, 2013, 14–25.

31. Oliveira, F.T.P., Aula, A., and Russell, D.M. Discriminating the relevance of web search results with measures of pupil size. *Proceedings of the 27th international conference on Human factors in computing systems*, ACM (2009), 2209–2212.

32. Onorati, F., Barbieri, R., Mauri, M., Russo, V., and Mainardi, L. Characterization of affective states by pupillary dynamics and autonomic correlates. *Frontiers in Neuroengineering 6*, (2013), 9.

33. Pan, B., Hembrooke, H.A., Gay, G.K., Granka, L.A., Feusner, M.K., and Newman, J.K. The determinants of web page viewing behavior: an eye-tracking study. *Proceedings of the 2004 symposium on Eye tracking research & applications*, ACM (2004), 147–154.

34. Platt, J.C. Advances in Kernel Methods. In B. Schölkopf, C.J.C. Burges and A.J. Smola, eds., *Advances in Kernel Methods - Support Vector Learning*. MIT Press, Cambridge, MA, USA, 1999, 185–208.

35. Pollatsek, A., Rayner, K., and Balota, D.A. Inferences about eye movement control from the perceptual span in reading. *Perception & Psychophysics 40*, 2 (1986), 123–130.

36. Preuschoff, K., Hart, B.M. 't, and Einhäuser, W. Pupil dilation signals surprise: evidence for noradrenaline's role in decision making. *Frontiers in Decision Neuroscience 5*, (2011), 115.

37. Rayner, K., Pollatsek, A., Ashby, J., and Jr, C.C. *Psychology of Reading*. Psychology Press, 2011.

38. Reichle, E.D., Pollatsek, A., and Rayner, K. E–Z Reader: A cognitive-control, serial-attention model of eye-movement behavior during reading. *Cognitive Systems Research 7*, 1 (2006), 4–22.

39. Vajjala, S. and Meurers, D. On The Applicability of Readability Models to Web Texts. *Proceedings of the Second Workshop on Predicting and Improving Text Readability for Target Reader Populations*, (2013), 59–68.

40. Villa, R. and Halvey, M. Is relevance hard work?: evaluating the effort of making relevant assessments. *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval*, ACM (2013), 765–768.

41. Voorhees, E.M. and Dang, H.T. Overview of the TREC 2005 Question Answering Track. *The XIV Text REtrieval Conference (TREC 2005) Proceedings*, (2005).

42. White, R.W. and Kelly, D. A Study on the Effects of Personalization and Task Information on Implicit Feedback Performance. *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, ACM (2006), 297–306.

43. Wickens, C.D. Multiple resources and performance prediction. *Theoretical Issues in Ergonomics Science 3*, 2 (2002), 159–177.

44. Wierda, S.M., Rijn, H. van, Taatgen, N.A., and Martens, S. Pupil dilation deconvolution reveals the dynamics of attention at high temporal resolution. *Proceedings of the National Academy of Sciences 109*, 22 (2012), 8456–8460.

45. Wilson, D. and Sperber, D. Relevance Theory. In G. Ward and L. Horn, eds., *Handbook of Pragmatics*. Blackwell, 2002.

46. Witten, I.H., Frank, E., and Hall, M.A. *Data Mining: Practical Machine Learning Tools and Techniques: Practical Machine Learning Tools and Techniques*. Elsevier, 2011.

47. Xie, B. and Salvendy, G. Prediction of Metal Workload in Single and Multiple Task Environments. *International Journal of Cognitive Ergonomics 4*, 3 (2000), 213–242.

48. Xu, J., Wang, Y., Chen, F., and Choi, E. Pupillary Response Based Cognitive Workload Measurement under Luminance Changes. In P. Campos, N. Graham, J. Jorge, N. Nunes, P. Palanque and M. Winckler, eds., *Human-Computer Interaction – INTERACT 2011*. Springer Berlin / Heidelberg, 2011, 178–185.

49. Xu, Y. (Calvin) and Chen, Z. Relevance judgment: What do information users consider beyond topicality? *Journal of the American Society for Information Science and Technology 57*, 7 (2006), 961–973.

50. Typeface anatomy. *Wikipedia, the free encyclopedia*, 2014. http://en.wikipedia.org/w/index.php?title=Typeface_anatomy&oldid=598703420.