The Effect of Threshold Priming and Need for Cognition on Relevance Calibration and Assessment

Falk Scholer
School of Computer Science
and Information Technology
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

Diane Kelly,
Wan-Ching Wu,
Hanseul S. Lee
School of Information and
Library Science
University of North Carolina
Chapel Hill, NC, USA
dianek@email.unc.edu

William Webber
College of Information Studies
University of Maryland
College Park, Maryland, USA
wew@umd.edu

ABSTRACT

Human assessments of document relevance are needed for the construction of test collections, for ad-hoc evaluation, and for training text classifiers. Showing documents to assessors in different orderings, however, may lead to different assessment outcomes. We examine the effect that *threshold priming*, seeing varying degrees of relevant documents, has on people's calibration of relevance. Participants judged the relevance of a prologue of documents containing highly relevant, moderately relevant, or non-relevant documents, followed by a common epilogue of documents of mixed relevance. We observe that participants exposed to only non-relevant documents in the prologue assigned significantly higher average relevance scores to prologue and epilogue documents than participants exposed to moderately or highly relevant documents in the prologue. We also examine how *need for cognition*, an individual difference measure of the extent to which a person enjoys engaging in effortful cognitive activity, impacts relevance assessments. High need for cognition participants had a significantly higher level of agreement with expert assessors than low need for cognition participants did. Our findings indicate that assessors should be exposed to documents from multiple relevance levels early in the judging process, in order to calibrate their relevance thresholds in a balanced way, and that individual difference measures might be a useful way to screen assessors.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and software—*performance evaluation*.

Keywords

Evaluation, assessors, relevance assessments, relevance behavior, need for cognition, threshold priming, order effects, calibration

SIGIR'13, July 28–August 1, 2013, Dublin, Ireland.

1. INTRODUCTION

Relevance is a fundamental concept in information retrieval (IR) and forms the basis of most methods of evaluating IR systems. Relevance assessments, however, are subject to human judgment. Different assessors may make different assessments of the relevance of the same document to the same topic, and one assessor may make different assessments of the same document at different times and under different conditions.

Relevance assessments are used for retrieval evaluation, either through reusable test collections such as those constructed by TREC and similar efforts, or through more ad-hoc evaluation. While differences in relevance assessments may have a limited effect on comparative system evaluation, they have a serious impact on the absolute evaluation of system effectiveness, as the level of agreement between human assessors places a practical upper bound on measurable system quality. This distortion is consequential in environments where absolute measures of retrieval completeness and accuracy are required, such as e-discovery, patent retrieval, and research literature surveys. Relevance assessments are also used to train text classifiers and learning to rank systems, and assessor variability may work to degrade the effectiveness of such systems.

While many causes of human variability in relevance assessment are inescapable, some are within the control of the evaluation designer [26]. One such factor is the order in which the assessor is asked to assess documents for relevance. A long sequence of irrelevant documents, for instance, might cause an assessor to lower their threshold of relevance, or alternatively to lose concentration and miss relevant documents—an effect evaluation designers could seek to counteract by seeding likely-relevant documents more evenly. The overall density of likely-relevant documents in the set of documents for assessment may be under the control of the evaluation designer, through sampling or pooling decisions. For instance, a uniform sampling of a document corpus will give a much lower density of relevant documents than a stratified sampling that weights samples towards retrieved or highly-ranked documents.

As assessors evaluate documents against a topic, whether of their own or of another's creation, they are building internal relevance models that guide their decision-making process about whether a document is relevant or not, and (where graded relevance is used) about the level of relevance to assign to documents. We posit that these resulting relevance models will develop differently depending on the relevance levels of documents that assessors encounter, and that this differential development of relevance models will result in different calibrations, or thresholds for relevance. We refer to this posited effect as *threshold priming*. For example, an assessor who encounters few or no relevant documents would (under our hypothesis) have a lower threshold for relevance than a person who encounters a large number of highly relevant documents. Subsequently shown the same documents, the former assessor would tend to assign a higher relevance grade than the latter.

A further factor influencing assessment reliability that may be under the evaluator's control is the choice of the assessors themselves. Studies in the e-discovery field have found that there is a great deal of variability in assessor reliability, but this does not seem to correlate with training or (at least legal) expertise (Section 2.2). Perhaps more fundamental personality traits and capacities are at play. One possible trait identified in the psychological literature is *need for cognition*, a measure of the extent to which individuals enjoy engaging in concentrated intellectual activity. We believe that assessing relevance is an example of concentrated intellectual activity, since it requires sustained attention, close and careful reading, and good memory. There are standardized scales for need for cognition; if this trait is correlated with assessor reliability, it could provide a useful filter for assessor recruitment.

We investigate the hypotheses that threshold priming and need for cognition influence assessor relevance judgments by asking the following three research questions:

1. Does threshold priming, based on the relevance level of documents that are encountered early in the judging process, affect how an assessor assigns relevance to documents that are seen later on?

2. Does threshold priming affect how well an assessor is able to form a stable conception of relevance for a topic, based on the self-consistency of ratings over time, and the level of agreement between assessors?

3. Does the need for cognition (NFC) characteristic of an assessor predict how they assign relevance to documents, or how long they take to make judgments?

2. BACKGROUND

We first review research from the information behavior literature about defining and measuring relevance. This research has primarily focused on the behaviors of searchers working on search tasks (both assigned and natural). Next, we review work examining how assessors make relevance judgments when engaged in the development of test collections or similar evaluation tasks. Finally, we review work describing individual differences and relevance behavior, and specifically need for cognition.

2.1 Relevance

A trilogy of articles by Saracevic [20, 21, 22] provide a comprehensive overview and synthesis of relevance research spanning over 30 years. Saracevic [21] conceptualized five classes of relevance: (1) system or algorithm; (2) topical; (3) pertinence or cognitive; (4) situational; and (5) motivational or affective. In our research, we focus on *topical relevance*, since this is the type of relevance that is modeled by the relevance assessments that accompany most test collections. *Topical relevance* is defined as "an intellectual assessment of how an information object corresponds to the topical area required and described by the request(s) for information" [2, p. 915]. We are also concerned with *cognitive relevance*, as our research is motivated by the theory that as people encounter documents, they are calibrating their internal relevance models of the topic, which serve as a basis for subsequent assessments. This is related to the idea of psychological relevance, which posits that relevance assessments are made in relation to a person's current psychological state: "...relevance judgments are a function of one's mental state at the time a reference is read. They are not fixed; they are dynamic" [12, p. 612]. Specifically, we are interested in how the order in which documents are encountered impacts people's relevance calibrations and subsequent assessments.

While it is generally accepted that order effects occur during document assessment, Saracevic [22] notes that only a few studies have systematically addressed this issue. The results of these studies have been inconsistent, likely due to variability in research methods. Eisenberg and Barry [9] experimented with two document orders for a set of 15 documents and a single query: one ranked high to low relevance, the other low to high. The authors found that in the high–low condition, people underestimated the relevance of documents at the higher end (assigning lower relevance scores to the highly relevant documents), while those in the low–high condition overestimated the relevance of documents in the low to middle range (assigning higher relevance scores to less relevant documents). This behavior was explained as a "hedging phenomenon" (p. 296), where participants working with a fixed scale and an unknown set of stimuli are reluctant to initially assign high or low scores to items. Purgaillis Parker and Johnson [19] did not find a systematic order effect in their study, but they used citations, rather than full-text, and like Eisenberg and Barry [9] only experimented with sets of 15 objects. Later, Huang and Wang [14] found that set size matters, with order effects being detected for sets of size 15, 30, 45 and 60, but not 5 and 75. They also found that relevance scores assigned by people in the low to high treatment were higher than those assigned by people in the high to low treatment, which is consistent with the results of Eisenberg and Barry [9].

In our research, we also focus on *degree of relevance*. Degree of relevance refers to the rating or indication of the relevance value given to documents. Borlund [2] provides an overview of different degrees of relevance including binary relevance, tripartite relevance, scale-based relevance and graded relevance. Borlund observes that degrees of relevance can also refer to whether the object as a whole is considered relevant, or only a part of it. Saracevic [22] indicates that previous research shows that people prefer to judge document relevance on a continuum, and comparatively. If people construct internal relevance models as they assess documents, then it seems reasonable to assume that the degree of relevance of the documents encountered will impact the formation of people's *relevance thresholds*, or the amount of evidence needed to associate documents with each degree of relevance.

In addition to Saracevic's conceptualizations, many researchers have also documented the various criteria (e.g., novelty, recency, depth) that people use when making relevance assessments [1, 25] and observed that the process of assigning relevance to documents is dynamic and inter-dependent, and impacted by a number of variables, including how much one knows about the topic and the search stage [6, 27]. There has also been a persistent body of research about how assessors make relevance judgments for test collections or other system evaluations. This work is reviewed in more detail in the next section.

2.2 Assessor Behavior

High levels of assessor disagreement about document relevance have been observed in a number of studies. Voorhees [29] reports a mean positive agreement between TREC assessors of 0.58. (Positive agreement can also be interpreted as the F_1 score that one assessor would achieve if evaluated by the other assessor.) Oard and Webber [18] survey a number of studies on assessor agreement, observing positive agreement between 0.33 and 0.76. Analyzing TREC assessment data, Scholer et al. [23] find that a single assessor will make at different times a different assessment of the binary relevance of the same document between 15% to 19% of the time, and 19% to 24% of the time for trinary relevance. Voorhees [29] finds that assessor disagreement has a limited effect on relative measures of system quality, but observes that it sets an upper bound on the practically measurable absolute quality. Carterette and Soboroff [5] find that (randomly simulated) optimistic assessors (those tending to mark more documents relevant) disrupt evaluation reliability more than pessimistic assessors do.

Webber [31] finds considerable variance between assessors' agreement with an authoritative assessor; nevertheless, Oard and Webber [18] conclude that assessor disagreement differs more between topics than between assessors, suggesting that topic difficulty is a major component in assessment variability. Grossman and Cormack [11] argue that assessor disagreement is due in 90% of cases to inarguable assessor error; however, their dataset has a strong selection bias towards such errors. Webber et al. [32] find that giving more detailed assessment guidelines does not improve assessor agreement or reliability. Webber et al. [32], Wang and Soergel [30], and Efthimiadis and Hotchkiss [8] all compare the reliability for e-discovery of legally trained assessors (lawyers or law students) with that of assessors who lack legal training, and find no difference between the two groups. Kazai et al. [16] find systematic biases between assessor groups in Web search towards, for instance, Wikipedia pages or documents rich in query keywords.

The impact of document order on relevance assessments has also been investigated in the context of crowd-sourcing. Le et al. [17] studied how the impact of the distribution of answers in a training set influences worker performance, and found that the accuracy of workers increases when the training set more closely reflects the underlying distribution of items that are matches or mismatches in a categorisation task. Kazai et al. [15] demonstrated that when workers are asked to assess a series of 10 documents ordered by their expected relevance (as derived from experimental IR systems) for an INEX book track, their overall accuracy compared to expert judgments is lower than when documents are ordered randomly.

Sormunen [24] investigates multi-level relevance criteria, carrying out extensive re-judging of documents for 38 topics from the TREC 7 and 8 newswire collections. Judgments were made on a 4-point relevance scale: highly relevant; relevant; marginally relevant; and non-relevant. The analysis shows that around half of the relevant documents were only marginally relevant, defined as not containing information beyond what is already available in the topic description. The assessments were made by students enrolled in a Masters degree program, and involved the pre-reading of topical documents before starting judgments. We make use of these *expert* judgments in this paper, using them as an underlying gold standard for our experiment.

2.3 Need for Cognition

While there has been some research about the impact of individual differences on search behavior (e.g., Ford et al. [10]), according to Saracevic [22] there has been little research exploring how individual differences impact relevance behavior. The exception is Davidson [7] who studied openness to information, which was measured by a number of cognitive style variables such as open-mindedness, rigidity and locus of control. Davidson [7] found that about 30% of the variance in relevance assessments was attributable to variances in openness to information.

In this study, we focus on one individual difference measure, *need for cognition* (NFC), which has been extensively studied in other fields as a way to understand differences in the extent to which people process information [3]. Cacioppo et al. [4] define NFC as "a stable individual difference in people's tendency to engage in and enjoy effortful cognitive activity" (p. 198). Empirical studies have found that individuals with high NFC are more motivated to process information, pay more attention to argument quality, perform better on cognitive tasks, react more positively to complex rules, generate more task-related and thoughtful responses, and recall more information about tasks (see Cacioppo et al. [4] for a detailed review). In this study, we seek to determine if and how NFC impacts relevance behavior.

3. METHODS

A between-subjects laboratory experiment with 82 participants was conducted. Each participant was given a single topic and asked to judge the relevance of 48 documents presented in a pre-specified order. The experimental sessions lasted approximately one hour. Participants were compensated with US$10.

3.1 Treatment

The treatment, or experimental manipulation, occurred within the lists of documents presented to participants. The document lists contained two parts: the *prologue*, consisting of the first 20 documents; and the *epilogue*, consisting of the last 28 documents. Documents in the prologue contained the experimental manipulations, while those in the epilogue were held constant for all participants for each topic. These manipulations and lists can be viewed in Table 1. Note that participants were presented with one document at a time, rather than a clickable list.

Within the prologue, one of three treatments was represented, *low*, *medium* or *high*, which corresponded to the degree of relevance of documents in the 'X' positions in the list in Table 1. In the low treatment, X was replaced by non-relevant documents. In the medium treatment, X was replaced by marginally relevant or relevant documents. In the high treatment, X was replaced by highly relevant documents. Documents shown in all positions other than those marked with an X were identical across treatments for each topic. The documents shown in positions 46-48 were duplicates of those shown in positions 21, 22 and 24. One document list was instantiated for each topic and treatment level combination; all participants who completed a specified treatment for a particular topic worked through the same list.

The documents, topics and relevance judgments came from a subset of the TREC-7 and TREC-8 test collections, which was created by Sormunen [24]. Marginally relevant and relevant documents were merged into one class and used in our treatments. However, participants were provided with the four-point relevance scale used in Sormunen [24] to make their judgments. The definitions of these various levels of relevance were:

- *Highly relevant (3)*: The document discusses the themes of the topic exhaustively. In case of a multi-faceted topic, all or most sub-themes or viewpoints are covered. Typical extent: several text paragraphs, at least 4 sentences or facts.
- *Relevant (2)*: The document contains more information than the topic description but the presentation is not exhaustive. In case of a multi-faceted topic, only some of the sub-themes or viewpoints are covered. Typical extent: one text paragraph, 2-3 sentences or facts.
- *Marginally relevant (1)*: The document only points to the topic. It does not contain more or other information than the topic description. Typical extent: one sentence or fact.
- *Not relevant (0)*: The document does not contain any information about the topic.

Prologue				Epilogue					
1	X	11	NR	21	MR-R	31	HR	41	MR-R
2	X	12	X	22	MR-R	32	MR-R	42	NR
3	NR	13	NR	23	NR	33	MR-R	43	MR-R
4	NR	14	X	24	MR-R	34	MR-R	44	NR
5	X	15	NR	25	MR-R	35	HR	45	HR
6	NR	16	X	26	HR	36	MR-R	46	MR-R, 21
7	X	17	X	27	NR	37	HR	47	MR-R, 22
8	NR	18	NR	28	HR	38	MR-R	48	MR-R, 24
9	X	19	NR	29	NR	39	NR		
10	NR	20	X	30	NR	40	HR		

Table 1: Design template for document lists. X indicates documents that differed among treatments in the Prologue. NR=non-relevant document; MR-R=relevant or marginally relevant document; HR=highly relevant document. Documents in positions 46-48 are duplicates of those in positions 21, 22, and 24.

When participants logged in to the system, they were presented with a description of what they would be doing during the experimental session. Participants were instructed: "The aim of this task is to assess the relevance of a set of documents for a particular search topic. You will be provided with a search topic to work on. You will then be presented with a series of documents, one at a time. Read each document, and decide if it is relevant for the topic." Participants were presented with the relevance definitions and scale, and were instructed to judge each document *independently* and *on its own merit* ("If a document contains information that makes it anything other than not relevant, then you should choose the appropriate relevance category, even if you have seen that information previously in another document."). Participants were further told that the documents were from the late 1980s and 1990s and that when a topic referred to "current" they should interpret this as "current to the time the article was written."

The assessment interface, which presented the search topic at the top of the screen (title, description and narrative), a document in the middle of the screen (one document was presented at a time), and the relevance assessment scale at the bottom of the screen, was also explained in the instructions. Participants were instructed that once they submitted an assessment they could not return to revise it later. During the assessment, participants could hover their mouse over each choice on the relevance scale to see the relevance definitions.

After participants finished reviewing the instructions, they completed a training task with a topic not used in the main study and two documents, so they could become familiar with the assessment process and interface. Several pilot tests were conducted during the development of the instructions and experimental infrastructure.

3.2 Topics

After completing the training task, participants were presented with one of three search topics. Our choice of topics was guided by several concerns. First, we needed to select topics for which there was a good assortment of documents at various relevance levels. Second, we needed to select topics about which most participants would know little, so that participants would start the study with the same basic models of the topics. Finally, we wanted to select topics that might interest participants. Several pilot tests were conducted during the topic and document selection phases.

Three topics were used in this study: 385 (hybrid fuel cars), 396 (sick building syndrome), and 415 (drugs, Golden Triangle). These topics were counter-balanced across the experimental treatments; that is, there were high, medium and low treatments for each topic. We understand that the inclusion of only three topics limits our ability to generalize, but including a larger number of topics was not possible since it would have required a prohibitively large number of participants to achieve statistical power.

	385	396	415
Frequency of past search	Never	Never	Never
Prior knowledge	A little	A little	Nothing
Interest	Slightly	Slightly	Somewhat
Relevance to life	Moderately	Slightly	Not at all

Table 2: Participants' Mode Responses to Topic Questions.

Participants were presented with the title, description and narrative fields and asked (1) How many times they had searched for information about the topic in the past (never, 1-2 times, 3-4 times, 5 or more times); (2) How much they knew about the topic (nothing, a little, some, a great deal); (3) How interested they were to learn more about the topic (not at all, slightly, somewhat, moderately, very); and (4) The relevance of the topic to their lives (not at all, slightly, somewhat, moderately, very). Following this, participants started the assessments.

Table 2 presents participants' mode responses to the topic items. Overall, most participants had not ever searched for information about these topics, knew little or nothing about them, were somewhat or slightly interested in learning more about them and found them moderately (385, hybrid fuel cars), slightly (396, sick building syndrome) and not at all (415, drugs, Golden Triangle) relevant to their lives. While we had hoped participants' interests in the topics might be slightly higher, our main objective was to select topics about which most participants would have no prior knowledge, so from this perspective, our topic selection was satisfactory.

3.3 Exit Questionnaire

After participants completed the main assessment task, they were directed to an exit questionnaire. The first question asked them to indicate what, if anything, was challenging about deciding which relevance levels to associate with each document. Next, they were asked to indicate their confidence in the relevance judgments they made (not at all confident, slightly confident, somewhat confident, moderately confident, very confident).

The next set of items consisted of the *Need for Cognition* scale (Cacioppo and Petty [3]). This 18-item scale contains statements such as "The notion of thinking abstractly is appealing to me," and "I like to have the responsibility of handling a situation that requires a lot of thinking." Participants indicated the extent to which the statements were characteristic of them using five choices (extremely uncharacteristic of me, somewhat uncharacteristic of me, uncertain, somewhat characteristic of me, and extremely characteristic of me).

The last part of the exit questionnaire contained demographic questions about participants' sex, age, student and occupational statuses, as well as whether English was their native language.

3.4 Participants

Participants were recruited through the mass email service at the University of North Carolina, Chapel Hill. A power analysis performed before the study indicated that approximately 25-30 participants were needed per treatment. Eighty-two people completed the experiment. Participants were randomly assigned to condition: 26 were in Treatment 1 (Low), 27 were in Treatment 2 (Medium) and 29 were in Treatment 3 (High). The majority of participants were female (85.37%). Their ages ranged from 18 to 55 years (M = 23.70, SD = 8.96). Seventy (85.36%) participants were students (57 undergraduates and 13 graduates); seven (8.54%) were university employees; four (4.88%) were both students and employees and one was neither. Student participants' majors were: social sciences (32%); professional schools (29%); sciences (19%); humanities (15%); and undecided (5%). For participants who were full- or part-time employees at the university (11, 13.42%), job titles included research assistant, IT support specialist, librarian, accountant, examiner and administrative manager/supporter. Seventy-one participants (86.59%) indicated they were native English speakers.

4. RESULTS

This paper investigates the effects of threshold priming and need for cognition on relevance assessments. In this section we present the results of our experiments as they relate to the three main research questions.

4.1 Relevance Assessments

The first research question aims to analyze how threshold priming, as operationalized by the relevance level of documents that participants see early in the judgment process, impacts how relevance scores are assigned to documents later on.

Recall that our experiments included three treatment conditions for documents in the prologue (the first 20 items that were judged): participants were shown either highly relevant (high), marginally relevant and relevant (medium), or non-relevant (low) documents. All participants then saw a consistent set of 28 documents in the epilogue. In this section, we investigate the impact of these threshold priming treatments in four ways: the impact on the relative assignment of relevance scores between study participants; the impact on relevance score assignments relative to underlying expert judgments; whether possible differences endure over time; and, whether the amount of time taken to make assessments differ according to treatment.

Before presenting these results, we first perform a treatment check to verify that participants in each of the three groups experienced the intended treatments. Participants' mean (standard deviation) ratings of prologue documents were 0.56 (0.40), 1.08 (0.34) and 1.42 (0.22) for each of the groups low, medium and high, so it appears that participants experienced the intended treatments. If participants were marking documents exactly as the underlying experts, we would expect those in the low group to have a mean of 0, those in the medium group to have a mean around 1 and those in the high group to have a mean of 1.5. These results show while participants experienced the intended treatments, those in the low group also up-marked some non-relevant documents. We explore this behavior in a subsequent section.

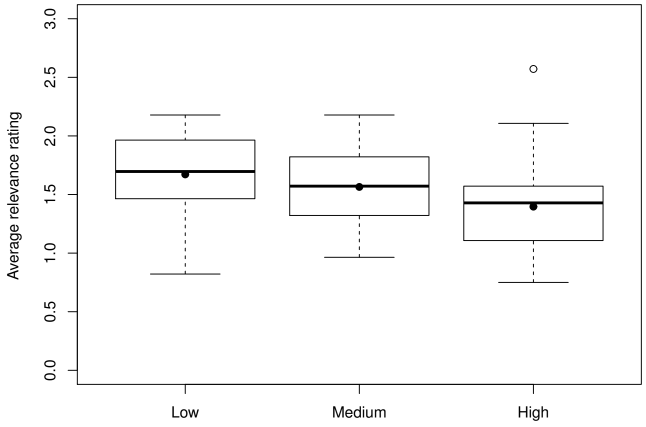

Figure 1: Distribution of mean relevance scores assigned to documents in the epilogue by participants in each of the three prologue treatments.

Impact on relative assignment of relevance scores. A convenient way to characterize the overall behavior of participants in the epilogue is to consider the mean relevance scores assigned to documents in positions 21–48. The statistic captures differences in how participants assigned relevance scores, and enables the analysis of *relative* judging behavior. (Since individual document assessments were made using a four-point ordinal scale, the mean scores should not be interpreted in absolute terms.)

A boxplot of these average relevance scores assigned by participants in each treatment is shown in Figure 1. The boxes show the data points of the 25th to 75th percentiles, with the solid black line representing the median, and mean values shown by solid black circles. The whiskers show the range of data and outlier values (data points that lie more than 1.5 times the inter-quartile range away from the box) are shown as circles. The means (and standard deviations, in parentheses) of the average relevance scores assigned to documents in the epilogue is 1.67 (0.38), 1.57 (0.33) and 1.40 (0.42) for the low, medium and high treatments, respectively.

Another possible source of variation that could impact the average relevance scores that were assigned in the epilogue comes from the search topics that participants were working on. The mean average relevance scores assigned by participants for each Topic was 1.57 (0.42) for Topic 385, 1.58 (0.37) for Topic 396, and 1.47 (0.39) for Topic 415.

A two-way ANOVA using type II sums of squares was conducted to investigate the statistical significance of both the treatment and topic effects. The treatment effect was found to be statistically significant ($F(2, 73) = 3.63, p = 0.031$) while the topic effect was not significant ($F(2, 73) = 0.61, p = 0.548$). The interaction effect was also not significant ($F(4, 73) = 0.64, p = 0.636$). Follow-up pairwise t-tests were conducted to further investigate the treatment effect (using the Bonferroni-Holm correction for multiple comparisons [13]). The results indicated that the difference between average scores assigned by participants in the low and high treatments differed significantly ($t(53) = 2.56, p = 0.025$). The differences between the other conditions were not significant ($p > 0.05$). In terms of relative behavior, it appears that participants who initially saw no relevant documents compensated by assigning higher relevance scores to items in the epilogue, while those who were exposed to highly relevant documents assigned more moderate scores.

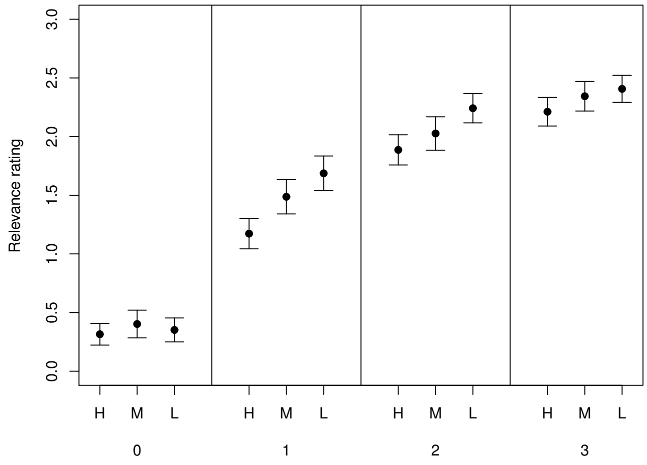

Figure 2: Mean relevance assessments of documents in the epilogue, grouped by the expert relevance rating of the document (0–3) and by prologue treatment.

Topic	Prologue			Epilogue		
	low	med	high	low	med	high
385	1.36	1.06	0.61	0.59	0.71	0.34
	0.15	*0.12*	*0.09*	*0.07*	*0.15*	*0.11*
396	0.31	0.12	0.26	0.29	0.30	0.40
	0.11	*0.05*	*0.09*	*0.11*	*0.11*	*0.21*
415	0.29	0.26	0.13	0.17	0.24	0.21
	0.06	*0.13*	*0.04*	*0.06*	*0.06*	*0.06*
Total	0.67	0.44	0.33	0.35	0.40	0.32
	0.12	*0.10*	*0.06*	*0.06*	*0.07*	*0.07*

Table 3: Mean of relevance assessments (standard error across assessors in italics) of common non-relevant prologue and epilogue documents, by topic and treatment.

Figure 2 shows the mean relevance assessments that participants assigned to documents (with a 95% confidence interval), grouped first by the expert relevance assessment and second by the experimental treatment. It can be seen that the treatment conditions had the strongest impact on ratings that participants assigned to the mid-range documents (that is, moderately relevant or relevant). These observations are supported by an analysis of the differences between the treatment conditions within each of the four expert relevance groups using a Kruskal-Wallis test ($p = 0.773$, $p < 0.001$, $p < 0.001$, and $p = 0.065$ for expert relevance levels 0 to 4, respectively).

The prologues contained ten non-relevant documents which were the same within each topic. Participants' assessments of these non-relevant documents can be seen in Table 3. The mean assessed relevances for these non-relevant documents were 0.65, 0.48, and 0.33 for the low, medium, and high treatments. A two-way ANOVA finds the difference between treatments statistically significant ($F(2, 73) = 9.78, p < 0.001$), with topic and topic–treatment interactions also being significant ($F(2, 73) = 68.01, p < 0.001$, and $F(4, 73) = 4.79, p = 0.002$). Post-hoc two-way ANOVAs between each pair of levels find significance for the contrasts of high with low treatment and medium with low treatment, but not high with medium treatment (respectively, $F(2, 49) = 20.03, p < 0.001$; $F(2, 50) = 6.91, p < 0.05$; and $F(2, 47) = 2.46, p > 0.05$). For the seven non-relevant documents in the epilogue, mean assessed relevances were 0.35, 0.42, and 0.32 for the low, medium, and high treatments, which was not significant in a two-way ANOVA, though there was still a significant topic effect ($F(2, 73) = 7.37, p = 0.001$). Assessors who saw only non-relevant documents in the prologue were disproportionately inclined to mark some of these documents as relevant; this effect disappeared in the epilogue, however, most likely due to having seen some genuinely relevant documents.

Impact on relevance scores in comparison to expert judgments. In addition to considering the impact that the treatments had on how participants in the different groups assigned relevance scores relative to each other, it is also possible to analyze how these scores compare to the underlying expert relevance judgments. As explained in Section 3, each epilogue consisted of an equal number of documents at each of the four relevance levels. The mean relevance level of each epilogue is therefore 1.5, by design. This expert mean score can be compared to the mean scores that were assigned under each of the treatment conditions; as indicated previously, these were 1.67 for low, 1.56 for medium, and 1.40 for high. Comparing each of these groups against the expert mean of 1.5 using a t-test shows that the scores of the low group differed significantly ($t(25) = 2.31, p = 0.0293$), while the medium and high groups did not show significant differences ($p = 0.354$ and $p = 0.193$, respectively).

A difference in the average rating that a participant assigns, relative to an expert rating, could arise due to a large difference in the assigned score for a small number of documents, or a smaller difference in the assigned score across a large set of documents. Figure 3 shows the frequency of the differences between the score that a participant assigned and the underlying expert judgment. All three treatment groups agreed with the expert judgments around 50% of the time. However, when differences occurred, the low group tended to assign relevance scores that were higher than that given by the experts, while the high group tended to assign scores that were lower than those given by the experts. Moreover, it can be seen that the observed differences in average scores were not simply due to the presence of a small number of items for which there were extreme differences of opinion.

The increase in average scores from low treatment participants on epilogue documents can be further analyzed by the expert relevance of each document. Due to bounding effects, there is a natural tendency for documents with low expert relevance to be assigned higher relevance by participants, and vice versa. We control for this by taking as the baseline the mean assessment of medium and high treatment participants for that document and topic. Compared to this baseline, low treatment participants on epilogue documents on average scored (expert-judged) irrelevant documents 0.01 points lower, marginally relevant documents 0.38 points higher, relevant documents 0.29 points higher, and highly-relevant documents 0.14 points higher. The difference between these score increases is significant in a one-way ANOVA test ($F(3, 724) = 8.47, p < 0.0001$). That is, the low treatment participants were not up-voting all documents in the epilogue, but only those with some evidence of relevance (as determined by the expert assessments). Moreover, they tended to boost the relevance of low-relevant documents by more than that of high-relevant documents.

Impact on assignment of relevance scores over time. While the prologue treatment conditions had an impact on the relative assignment of relevance scores to documents in the epilogue,

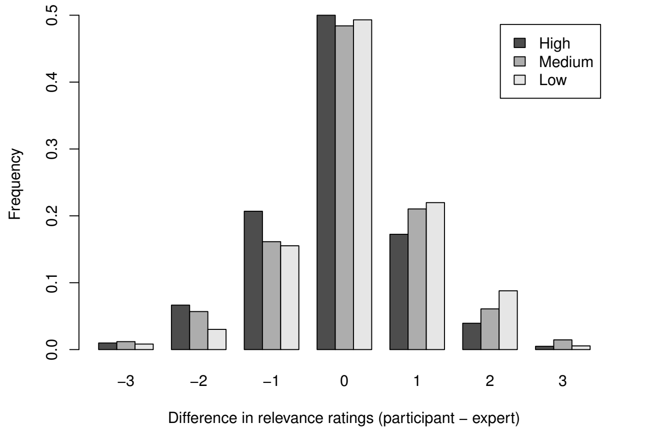

Figure 3: Frequency of the difference in relevance scores assigned by participants and experts to all epilogue documents.

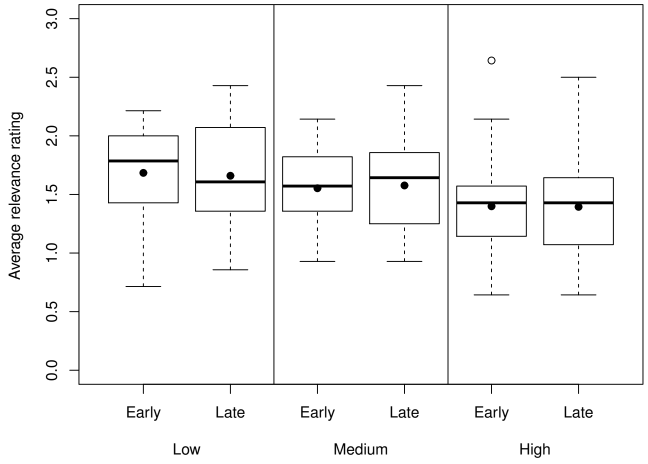

Figure 4: Distribution of mean per-participant relevance scores assigned to documents in the epilogue, for each of the three treatments.

this raises the question of whether such an effect was enduring over the entire 28 epilogue documents, or whether it changed over time. To investigate this, the epilogue was divided into halves, giving a group of 14 *early* documents and 14 *late* documents.

The results of this time-based split are shown in Figure 4, where the low, medium and high treatment conditions have been further partitioned into early and late groups. The graph suggests that variability (based on the inter-quartile range) is higher for the late documents than for the early documents. However, the differences between early and late relevance assignments within a particular treatment group are not statistically significant (t-test, $p > 0.1$).

This suggests that participants continue to refine their mental relevance models over time. Even if they do not have any reference points to begin with, they are able to re-calibrate once they begin to see documents that are relevant to different degrees.

Time taken to judge relevance. It is possible that the occurrence of documents with different levels of relevance has an effect on the amount of attention—and in particular, time—that participants devote to examining documents. For example, if an individual reads a number of documents and they are all non-relevant, the person might become disheartened, and as a result pay less attention to subsequent documents that are presented.

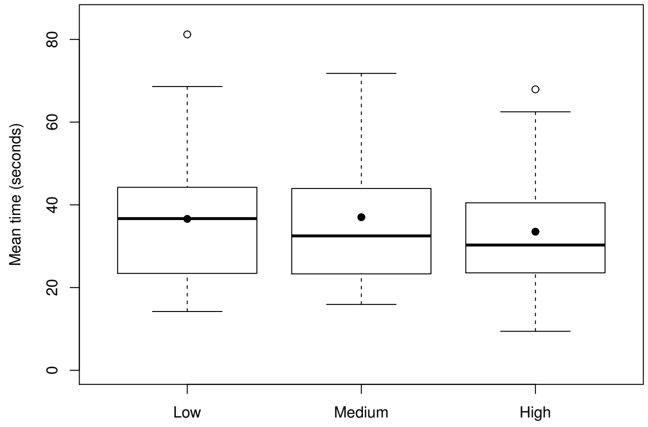

Figure 5: Distribution of the mean time that participants took to judge the relevance of documents in the epilogue, for each of the three treatments.

The boxplot in Figure 5 shows the mean time in seconds that participants took to make relevance judgments for the 28 documents in the epilogue, split by the three treatment conditions. The time was measured in seconds, from when the document was first displayed, until the participant entered and saved a response in the judging interface. While the mean judging times show slight variation (36.60, 37.02, and 33.51 seconds for the low, medium and high groups respectively), these differences are not statistically significant (one-way ANOVA, $F(2, 79) = 0.40, p = 0.669$).

4.2 Agreement

The second research question focuses on whether threshold priming affects how well participants are able to form a stable conception of relevance for a topic, based on the self-consistency of ratings over time, and the level of agreement between participants. Two types of agreement are considered: agreement among participants; and participant self-agreement when making repeat judgments of the same documents at different points in time.

Overall agreement among participants. Our experimental framework was constructed with reference to multi-level relevance judgments created by Sormunen [24], which were created through a careful assessment process.

Consider two sets of relevance judgments made by different assessors. The *overlap* (or percentage agreement) between these judgments is defined as the intersection divided by the union of the two sets (that is, the number of documents that were given the same relevance score by both assessors, divided by the total number of documents that were assessed). This mean pairwise percentage agreement among all participants is 44.80%.

When people read documents in response to a topic, their understanding of the topic changes. It is possible that exposure to different documents will influence their conception of relevance. For example, an individual who sees many relevant documents early might be able to more quickly develop a model of topical relevance; conversely, a person who has mostly seen non-relevant documents might find it more challenging to establish a stable model. To explore this issue, we investigate the relationship between study treatment and the level of agreement between participants when judging documents in the epilogue.

The treatment groups show only minor differences in mean pairwise agreement levels, with 46.45% (12.62), 45.28% (10.17) and 44.25% (12.85) for the low, medium and high groups, respectively. A two-way ANOVA was conducted to investigate the significance of treatment and topic effects. The results show that the treatment main effect was not statistically significant ($F(2, 326) = 0.99, p = 0.373$). However, differences in agreement due to the topic main effect were significant ($F(2, 326) = 5.70, p = 0.004$); the corresponding mean pairwise agreement levels were 42.76 (12.33), 44.79 (12.44), and 47.99 (10.65) for topics 385, 396, and 415, respectively.

This finding of a significant topic effect is consistent with other studies [18]. The interaction between topic and treatment was not significant.

Self-agreement over time. A second perspective on agreement is whether participants agree with their own relevance assessments, over time. Recall that the epilogue was constructed so that the first three "medium level" documents that a participant encountered in the epilogue recurred as the final three documents in the list. Self-agreement is calculated as the overlap between the relevance ratings assigned to these documents.

The average self-agreement across participants was 51.62%. Based on the different treatment levels, only slight variations were introduced, with self-agreement of 52.56%, 49.38% and 52.87% for the low, medium and high groups. The differences between the groups are not statistically significant, based on a one-way ANOVA ($F(2, 79) = 0.09, p = 0.918$).

4.3 Need for Cognition

Our third research question was whether need for cognition (NFC) influences relevance judgements. In this section, we consider the relationship between NFC and the relevance ratings participants assigned to the epilogue documents; the time taken to make relevance judgements; and the level of agreement between participants and experts.

To examine the effect of NFC, participants' responses to each of the 18 items on the NFC scale were averaged to arrive at a composite NFC score for each participant. Their composite scores from the scale ranged from 1.84 to 4.37. The mean and median of NFC composite scores were 3.16 and 3.24, respectively, and the standard deviation was .56. Participants were divided into a high need for cognition group (HNFC, n=41) and a low need for cognition group (LNFC, n=41) based on a median split [28]. The distribution of HNFC and LNFC participants across treatments was not statistically different ($\chi^2(2) = 2.741, p = 0.254$).

Impact on assignment of relevance scores. Overall, LNFC participants tended to assign lower relevance scores ($M = 1.49$, $SD = 0.39$) than HNFC ($M = 1.58$, $SD = 0.39$). A two-way ANOVA (treatment x NFC) was used to investigate the potential main and interaction effects of NFC and treatment on the relevance scores of documents in the epilogue (Figure 6). Results showed a significant main effect for treatment ($F(2, 76) = 3.89, p = 0.025$), but not NFC group ($F(1, 76) = 1.118, p = 0.294$). As the *prologue* treatment varied from *low* to *high*, the difference in relevance scores between HNFC participants and LNFC began to diverge, but not significantly ($F(2, 76) = 0.267, p = 0.767$).

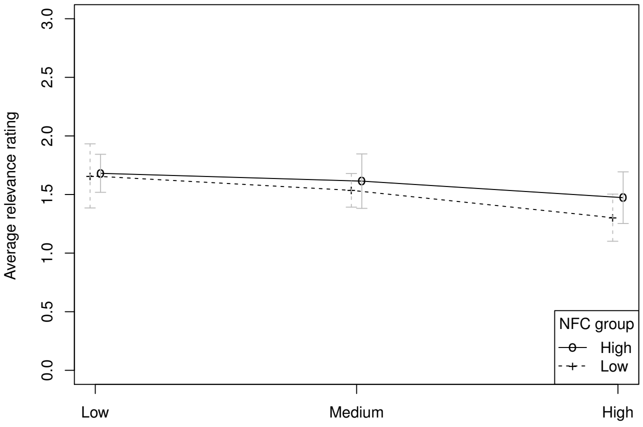

Figure 6: Main and interaction effects of NFC and prologue treatments on mean relevance scores assigned to epilogue documents.

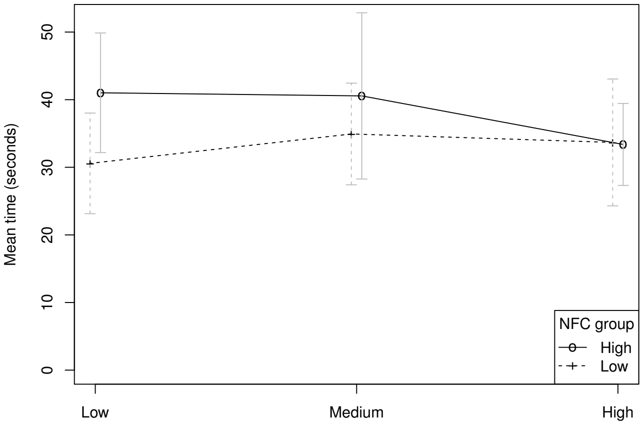

Figure 7: Main and interaction effects of NFC and Prologue treatments on mean time taken to judge the relevance of documents in the epilogue.

Impact on time to judge relevance. Results showed that HNFC participants spent more time making relevance judgements ($M = 37.97$, $SD = 16.33$) than LNFC participants ($M = 33.36$, $SD = 15.24$) (Figure 7). Results also showed that as treatment varied from *low* to *high*, the differences between the time taken by participants in the HNFC and LNFC groups converged. A two-way ANOVA found no significant differences (Treatment: $F(2, 76) = 0.473, p = 0.625$; NFC: $F(1, 76) = 2.154, p = 0.146$; Interaction: $F(2, 76) = .772, p = 0.466$).

Impact on agreement with expert judgements. Finally, the effect of NFC on participant agreement with the expert assessors was examined. Agreement was measured as the proportion of documents in the epilogue to which participant and expert assessor gave the same relevance value. HNFC participants had higher agreement ($M = 51.8\%$, $SD = 10.2\%$) with the experts than LNFC participants ($M = 46.7\%$, $SD = 9.0\%$). A three-way ANOVA (treatment x NFC group x topic) was performed to test this difference for significance. (Topic was included in the analysis due to the strong topic effect on agreement between assessors noted in Section 4.2.) The NFC effect was found to be significant ($F(1, 70) = 5.48, p = 0.022$); none of the other effects or interactions achieved significance.

4.4 Assessment Challenges

At the end of the study, participants were asked what, if anything, they found challenging about deciding which relevance levels to associate with documents. The most common challenge identified by participants was related to the extent to which the topic was represented in the document. Participants noted that they struggled with documents that contained relevant terms, but no real discussion of the issues. Some mentioned difficulties assessing documents that only mentioned a single aspect of the topic and those where the topic was not the main focus. Others struggled to assess documents that only contained a few sentences about the topic. Specifically, participants commented that the proportion of the document "about" the topic was something they had a difficult time dealing with when making relevance assessments. Another frequently mentioned challenge was document length, which according to many participants made skimming difficult and also placed a burden on their memory as they moved through the document trying to identify and track relevant information. Only two participants commented about fatigue.

Although it is generally assumed that assessors will (and can) base their judgements on topical relevance alone, many participants' comments were related to other types of relevance. With respect to cognitive relevance, participants indicated that their lack of background knowledge made the assessment task challenging. One participant commented that if he did not understand a document, he categorized it as not relevant or marginally relevant. Another participant observed that it was difficult to ignore novelty when making assessments. With respect to situational relevance, several participants indicated a desire for a more thorough topic description including information about why the information was needed and how it would be used. Finally, several participants commented that they did not like some of the document genres and formats because the display was unappealing, and this was difficult to ignore when making assessments (affective relevance). Overall, these comments challenge the notion that in assessment situations, relevance judgements are based purely on external representations (i.e., topic description) and that other types of relevance can be controlled.

5. CONCLUSION AND FUTURE WORK

This study investigated threshold priming, or the extent to which the relevance of documents viewed early during the assessment process impacted subsequent assessments. This study also investigated how need for cognition, an individual difference measure, impacted relevance assessments.

Our first research question examined how threshold priming impacted participants' relevance scores. We found that participants in the low treatment group assigned significantly higher mean relevance scores to documents in the epilogue than participants in the high treatment group. In particular this change in behavior was significant for documents that were in the middle of the relevance range (marginally relevant or relevant). To investigate the impact on relevance scores over time, we sub-divided documents in the epilogue into two halves (early and late), but found that mean relevance ratings did not differ significantly. In comparing participants' assessments of identical non-relevant documents in the prologue, we found a significant difference in the scores, with those in the low treatment group assigning the highest scores, followed by those in the medium and high treatment groups. This difference disappeared in the epilogue. Taken together, these results provide evidence that people's internal relevance models are impacted by the relevance of the documents they initially view and that they can re-calibrate these models as they encounter documents with more diverse relevance scores.

With respect to agreement among scores assigned by participants and the underlying expert relevance scores, we found that scores assigned by participants in the low group significantly differed from the expert judgments, while those assigned by participants in the medium and high groups did not. In looking more closely at the disagreements, we found that participants in the high group tended to assign scores that were lower than those given by experts, while participants in the low group tended to assign scores that were higher than those given by experts. In particular, participants in the low group tended to boost the relevance of low-relevant documents more than high-relevant documents, which is consistent with past research [9, 14]. These participants were not up-voting all documents in the epilogue, but rather only those with some evidence of relevance. Given that these participants also assigned higher ratings to non-relevant documents in the prologue, it is likely that they developed lower relevance thresholds for at least some categories of relevance.

Our second research question focused on whether threshold priming impacted how well participants were able to form a consistent conception of relevance for a topic. We found no significant differences in self-agreement levels, agreement between assessors, or agreement with experts, according to treatment condition. However, regardless of treatment condition, a reasonable level of disagreement remains, indicating that making relevance assessments is a challenging task and subject to substantial variation. An interesting question for future work is to investigate whether the low self-agreement for individual assessors over time is due to a genuine change in their internal relevance model, or is representative of general inherent variability in such models, perhaps due to factors such as mental fatigue.

Our final research question focused on whether *need for cognition* (NFC) influenced relevance judgments. Although high NFC participants assigned higher relevance scores to documents, this was not significant. We also observed a divergence in mean relevance score assigned by high and low NFC participants as the treatment groups varied from low to medium to high, but this also was not significant. Similar results were found for time taken to judge relevance. Overall, high NFC participants spent more time assessing documents, and the time spent by high and low NFC participants converged as treatments varied from low to medium to high. Finally, we found that high NFC participants' level of agreement with the *expert* assessors was significantly higher than that of low NFC participants. While not all of these results are statistically significant, the results suggest that individual difference measures might provide insight into assessor variability and could be a potentially useful way to screen assessors for characteristics that are associated with more consistent judgments.

Although we did not pose a research question about the challenges participants experience when making assessments, we included a question at the close of the study about this issue. Common challenges identified by participants concerned the depth of treatment of the topic, focus of the document, and the proportion of the document devoted to the topic. We observed many comments that indicated participants struggled to base their assessments purely on topical relevance, and instead wanted to consider cognitive, situational and affective relevance as well. These results show the difficulty in trying to restrict human judgments to topical relevance.

We are conducting a follow-up study using the same design and infrastructure except that participants are asked to justify their relevance assessments explicitly after viewing each document. This will allow us to elicit and monitor the formation of participants' internal relevance models and evaluate how this procedure impacts relevance assessments. Since the study design is the same, we will be able to compare the judgments of these participants with a randomly selected subset of participants from the current study to determine how this reflective procedure impacts the judgment process across experimental treatments. If this procedure results in more consistency in judgments across treatments, then it might be included as part of assessors' initial training.

Acknowledgements

This work was supported in part by the ICT Centre of Excellence program, National ICT Australia (NICTA), and in part under the Australian Research Council's Discovery Projects funding scheme (project number DP130104007). This material is based in part upon work supported by the National Science Foundation under Grant No. 1065250. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

[1] Carol Barry. User-defined relevance criteria: an exploratory study. *Journal of the American Society for Information Science*, 45: 149–159, 1994.

[2] Pia Borlund. The concept of relevance in IR. *Journal of the American Society for Information Science and Technology*, 54(10): 913–925, 2003.

[3] John T. Cacioppo and Richard E. Petty. The need for cognition. *Journal of Personality and Social Psychology*, 42(1):116–131, 1982.

[4] John T. Cacioppo, Richard E. Petty, Jeffery A. Feinstein, and W. Blaire G. Jarvis. Dispositional differences in cognitive motivation: The life and times of individuals varying in need for cognition. *Psychological Bulletin*, 119(2):197–253, 1996.

[5] Ben Carterette and Ian Soboroff. The effect of assessor errors on IR system evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 539–546, Geneva, Switzerland, 2010.

[6] Carlos A. Cuadra and Robert V. Katter. Opening the black box of relevance. *Journal of Documentation*, 23(4):291–303, 1967.

[7] David Davidson. The effect of individual differences of cognitive style on judgments of document relevance. *Journal of the American Society for Information Science*, 28(5):273–284, 1977.

[8] Efthimis N. Efthimiadis and Mary A. Hotchkiss. Legal discovery: does domain expertise matter? *Proceedings of the American Society for Information Science and Technology*, 45:1–2, 2008.

[9] Michael Eisenberg and Carol Barry. Order effects: a study of the possible influence of presentation order on user judgments of document relevance. *Journal of the American Society for Information Science*, 39:293–300, 1988.

[10] Nigel Ford, David Miller, and Nicola Moss. Web search strategies and human individual differences: cognitive and demographic factors, internet attitudes and approaches. *Journal of the American Society for Information Science and Technology*, 56:741–756, 2005.

[11] Maura Grossman and Gordon V. Cormack. Inconsistent assessment of responsiveness in e-discovery: difference of opinion or human error? In *DESI IV: The ICAIL Workshop on Setting Standards for Searching Electronically Stored Information in Discovery Proceedings*, pages 1–11, June 2011.

[12] Stephen P. Harter. Psychological relevance and information science. *Journal of the American Society for Information Science*, 43: 602–615, 1992.

[13] Sture Holm. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, 6(2):65–70, 1979.

[14] Mu-hsuan Huang and Hui-yu Wang. The influence of document presentation order and number of documents judged on users' judgments of relevance. *Journal of the American Society for Information Science and Technology*, 55(11):970–979, 2004.

[15] Gabriella Kazai, Jaap Kamps, Marijn Koolen, and Natasa Milic-Frayling. Crowdsourcing for book search evaluation: impact of hit design on comparative system ranking. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 205–214, Beijing, China, 2011. ACM.

[16] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and S. M. M. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 105–114, Maui, Hawaii, USA, 2012.

[17] John Le, Andy Edmonds, Vaughn Hester, and Lukas Biewald. Ensuring quality in crowdsourced search relevance evaluation: The effects of training question distribution. In *SIGIR 2010 workshop on crowdsourcing for search evaluation*, pages 21–26, 2010.

[18] Douglas W. Oard and William Webber. Information retrieval for e-discovery. 2013. In submission. http://ediscovery.umiacs.umd.edu/pub/ow12fntir.pdf.

[19] Lorraine M. Purgaillis Parker and Robert E. Johnson. Does order of presentation affect users' judgment of documents? *Journal of the American Society for Information Science*, 41(7):493–494, 1990.

[20] Tefko Saracevic. Relevance: a review of and a framework for the thinking on the notion of information science. *Journal of the American Society for Information Science*, 26(6):321–343, 1975.

[21] Tefko Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. part II: Nature and manifestations of relevance. *Journal of the American Society for Information Science and Technology*, 58:1915–1933, 2007.

[22] Tefko Saracevic. Relevance: a review of the literature and a framework for thinking on the notion in information science. part III: Behavior and effects of relevance. *Journal of the American Society for Information Science and Technology*, 58:2126–2144, 2007.

[23] Falk Scholer, Andrew Turpin, and Mark Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 1063–1072, Beijing, China, 2011.

[24] Eero Sormunen. Liberal relevance criteria of TREC – counting on negligible documents? In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 324–330, Tampere, Finland, 2002.

[25] Anastasios Tombros, Ian Ruthven, and Joemon M. Jose. How users assess web pages for information seeking. *Journal of the American Society for Information Science and Technology*, 56(4):327–344, 2005.

[26] Christian Unkelbach, Vanessa Ostheimer, Frowin Fasold, and Daniel Memmert. A calibration explanation of serial position effects in evaluative judgments. *Organizational Behavior and Human Decision Processes*, 119:103–113, 2012.

[27] Pertti Vakkari. Changes in search tactics and relevance judgments when preparing a research proposal: a summary of the findings of a longitudinal study. *Information Retrieval*, 4(3):295–310, 2001.

[28] Bas Verplanken. Need for cognition and external information search: Responses to time pressure during decision-making. *Journal of Research in Personality*, 27(3):238–252, 1993.

[29] Ellen M. Voorhees. Variations in relevance judgements and the measurement of retrieval effectiveness. *Information Processing and Management*, 36(5):697–716, 2000.

[30] Jianqiang Wang and Dagobert Soergel. A user study of relevance judgments for e-discovery. *Proceedings of the American Society for Information Science and Technology*, 47:1–10, 2010.

[31] William Webber. Re-examining the effectiveness of manual review. In *Proc. SIGIR Information Retrieval for E-Discovery Workshop*, pages 2:1–8, Beijing, China, July 2011.

[32] William Webber, Bryan Toth, and Marjorie Desamito. Effect of written instructions on assessor agreement. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 1053–1054, Portland, Oregon, USA, 2012.