Wakeling, Simon and Halvey, Martin and Villa, Robert and Hasler, Laura (2016) A comparison of primary and secondary relevance judgements for real-life topics. In: Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval (CHIIR '16). ACM, New York, pp. 173-182. ISBN 9781450337519 , http://dx.doi.org/10.1145/2854946.2854968

A Comparison of Primary and Secondary Relevance Judgements for Real-Life Topics

Simon Wakeling[1] Martin Halvey[2] Robert Villa[1] Laura Hasler[1]

[1] University of Sheffield
Information School, Regent Court,
Sheffield, United Kingdom.

[1] {first initial.surname}@sheffield.ac.uk

[2] University of Strathclyde,
Department of Computer and Information Sciences,
Glasgow, Scotland, United Kingdom

[2] martin.halvey@strath.ac.uk

ABSTRACT
The notion of relevance is fundamental to the field of Information Retrieval. Within the field a generally accepted conception of relevance as inherently subjective has emerged, with an individual's assessment of relevance influenced by numerous contextual factors. In this paper we present a user study that examines in detail the differences between primary and secondary assessors on a set of "real-world" topics which were gathered specifically for the work. By gathering topics which are representative of the staff and students at a major university, at a particular point in time, we aim to explore differences between primary and secondary relevance judgements for real-life search tasks. Findings suggest that while secondary assessors may find the assessment task challenging in various ways (they generally possess less interest and knowledge in secondary topics and take longer to assess documents), agreement between primary and secondary assessors is high.

CCS Concepts
• Information systems~Test collections • Information systems~Relevance assessment

Keywords
Assessment; Secondary; Primary; Judgement; Test Collection.

1. INTRODUCTION
The notion of relevance is central to Information Science research and there exists a vast body of literature on the subject. Whilst research into the notion of relevance is on-going, with many perspectives and open/unanswered questions, a generally accepted conception of relevance as inherently subjective has emerged, with an individual's assessment of relevance influenced by a myriad of contextual factors [16]. In particular understanding and measuring relevance is of the utmost importance to the evaluation of information retrieval (IR) systems. There are many views on evaluation with the IR community, where user-focused evaluation and system focused evaluation can be considered as two extreme points on a continuum of IR evaluation [11], with many points in between. An integral component of the system-orientated evaluation process is the generation of annotated test-collections. There are many successful initiatives in this area including TREC, CLEF, INEX etc. While processes vary between test collections, the method of creating a test collection typically consists of expert assessors creating a series of topics against which to assess the relevance of documents in a collection. In recent years there has been an increasing focus on this aspect of the evaluation process, particularly in terms of the accuracy and efficiency of the judgement process. Investigating ways in which this type of system focussed evaluation can be improved is an on-going effort, and has been the subject of much debate (e.g. [1]) and long-term critique, especially concerning the lack of user interactivity [15].

CHIIR '16, March 13-17, 2016, Carrboro, NC, USA © 2016 ACM.
ISBN 978-1-4503-3751-9/16/03...$15.00
DOI: http://dx.doi.org/10.1145/2854946.2854968

At the other end of the spectrum, from a user centred perspective, work lead by Borlund has noted the importance of task creation in the evaluation process, in particular that "an information need ought to be treated as a user-individual and potentially dynamic concept" [6]. In proposing the simulated work task situation as a method of stimulating that information need, Borlund argues that measures of system interaction and judgements of situational relevance better reflect real-life when the information need is truly realistic and engaging. She argues that this approach can lead to a more effective evaluation of a system.

While the topics created for TREC style relevance assessments may be user-individual, we argue that they do not in practice always represent "real-life" information needs. While a growing body of research has examined differences between primary and secondary relevance judgements for these synthetic tasks [2; 17], little has been done to investigate how the use of real-life search tasks might affect relevance judgements for test-collections. This paper aims to address this deficit by exploring the differences between primary and secondary relevance judgements for real life non-synthetic tasks gathered from staff and students at a large University. More specifically, we seek to answer the following research questions in relation to real-life search-tasks:

RQ 1: How does relevance assessment behaviour differ between primary and secondary assessors?
RQ 2: To what extent do secondary assessments agree with primary judgements?
RQ 3: To what extent do interest in and knowledge of the topic affect relevance judgements?
RQ 4: Does the length of the topic description affect secondary relevance judgements?
RQ 5: How does confidence in judgements differ between primary and secondary assessors?

To address these questions a two-part study was conducted where participants' real world information needs were gathered, and used

to generate document sets. These were then assessed by both the initial participant and a number of secondary assessors. A mixed methods approach was taken, where quantitative data relating to the judgement and assessment process was integrated with qualitative data collected during post-task interviews.

The rest of this paper is structured as follows: first, a short literature review of the recent and relevant material is provided, followed by a description of the study carried out. This includes a description of the data collected, with an emphasis on the qualitative data collected. Results are provided in Section 4, which is followed by a discussion of our findings and finally our conclusion.

2. RELATED WORK

2.1 Primary vs. Secondary Assessments

Voorhees [20] used TREC data to compare differences between the original relevance judgements of topic authors, and subsequent secondary relevance assessments. While significant variation in relevance judgements were observed, these were found to not meaningfully effect the subsequent evaluation of retrieval performance. Webber & Pickens [21] examined disagreements between primary and secondary assessors of a text classifier, finding that while the use of secondary assessors lowered the classification quality this had little practical effect on results rankings. Alonso & Mizzaro [3] examined agreement between TREC assessors and crowd workers on Mechanical Turk. They found that while agreement varied for individual assessors, collective agreement levels were high. Al-Harbi and Smucker [2] also compared the original judgements of TREC assessors with those of secondary assessors, focusing particularly on the assessment process. Using a think-aloud protocol they identified three general reasons for disagreements between primary and secondary assessments; *topic* (the secondary assessor having difficulty understanding or applying the topic description to the documents), *document* (difficulty processing the document), and *assessor* (the secondary assessor lacking knowledge or concentration).

It should be noted that while the primary assessments used in [20; 21] represent the judgement of the "creators" of a topic, these topics do not necessary represent real-life information needs. Chouldechova & Mease [7] in contrast, compared the relevance judgements of primary and secondary assessors for results sets returned by real-life search engine queries. They found that using primary assessments led to more valuable relevance judgements, attributing this to the superior background knowledge of primary assessors. The authors do however note some of the practical difficulties of eliciting such real-life primary judgements.

2.2 Impact of Domain Knowledge

Bailey et al. [4], Kinney et al. [12], Ruthven et al. [14] all found evidence that an assessor's level of topic knowledge positively correlated with judgement quality, with [14] also finding that interest in the topic was similarly related. Clough et al. [8] compared crowdsourced relevance judgements of search engine rankings with expert assessments, and concluded that while overall rankings were comparable, experts were able to better distinguish between different levels of highly accurate results. In contrast to these studies, Efthimiadis and Hotchkiss [9] compared expert and non-expert judgements for search topics within the TREC legal track, finding that the judgements of assessors without legal expertise were of higher quality than those of experts. Research has also suggested a link between domain knowledge and confidence in a judgement. Ruthven et al. [14] found that an assessor's prior confidence in their ability to judge documents for a topic was linked to their knowledge of the topic, and that this confidence level was found to influence their judgements. Al-Harbi and Smucker [2] suggest that secondary assessors are frequently uncertain about their judgement, which in extreme cases results in the assessor decision is being a guess. They advocate the collection of a certainty measure with each relevance judgement.

2.3 Impact of Topic Description

Al-Harbi and Smucker [2] found some evidence that the length of the topic description influenced differences in judgement between primary and secondary assessors. In particular they suggest that a short topic description may encourage a higher number of relevant judgements from secondary assessors, who are able to interpret the criteria for relevance more liberally. This contrasts with the work of Webber et al. [22], who found that in the context of the TREC legal track more detailed descriptions did not improve assessor reliability. It is also useful to note that reviews of variations in topic description structure and length across TREC programmes reveal an acknowledgement of the influence descriptions have on the judgement process. We note for example that TREC-4 shortened the length of descriptions, and removed the "narrative" section, which was found to greatly impact performance [10].

3. USER STUDY

3.1 Overview

The overall aim of the study was to explore differences between primary and secondary relevance judgements for real-life search tasks. As such the study was split into two parts: an initial questionnaire to gather real-life "search tasks" from participants, and an in-lab study which involved participants judging the relevance of documents to both their own and other participants' tasks. The first part was based on library search forms and the procedure is described in Section 3.1. From these search forms a document collection was generated (Section 3.2). These documents were then utilised in a lab study to gather assessments (Section 3.3 and 3.4).

While full details of the study design are below, several key decisions were made early in the study design that merit discussion here. Since it was necessary to generate documents relating to real-world search tasks, which would naturally cover a diverse range of topics, the web was used as a source for documents. It was also decided to elicit specific types of search-task from respondents to the phase one survey. The vast literature on information seeking has resulted in a variety of ways of categorising search-tasks, but the evaluation of the impact of all possible task types is beyond the scope of the research presented here. For the purposes of this study it was deemed sufficient to explicitly distinguish between two fundamental types of task – open (a task in which the searcher will likely need to access and synthesise information from several sources to address their information need, and for which there may not be a single definitive answer) and closed (a task which likely has a single unambiguous solution) [13].

The structure of the topic description was modelled on early TREC protocols [10], and consisted of three sections: a basic description of the topic, an outline of the context for the search, and an explicit summary of the criteria for assessing relevance. The assessment itself took the form of a binary relevant/not relevant judgement. Since results presented in related studies (e.g. [2; 17]) are also based on binary judgements, we determined that the use of a scale or continuum would potentially affect the comparability of our results. A binary judgement was therefore collected for each document. Finally the collection of qualitative data was done

through post-session play-back interviews rather than the think-aloud protocol used by [2]. This was to allow for the collection of temporal and behavioural data relating to the judgement process, using a think aloud protocol could have potentially skewed the data collected.

3.2 Task Generation

Participants were initially recruited via an introductory email sent to volunteer mailing lists at the University of Sheffield. This email explained what would be required of participants, and offered compensation of £24 for those completing both stages of the research project. Those interested in participating were asked to email the investigators directly, and the first 20 respondents were then sent two links; to an online calendar to book a date and time for the lab session, and to an online task form. This form first gathered some background information about participants (age, gender, educational background etc.), and then asked for details of two search tasks the participant either was about to undertake, or had recently undertaken. The form specified that the first task should be a closed search task, and the second an open search task. Explanations of both terms and example search tasks were provided to ensure participants understood what was required. To elicit details of each search task, participants were asked to respond to four requests for information about the search task. These are presented below, along with an example response from a participant:

1. Please describe what you are searching for, in one clear and precise sentence.
   What led to the recession that began in 2008?
2. Please describe your search situation in more detail (e.g. the context of your search, the purpose of seeking this information, why you are interested in it, etc.) A good way of approaching this is to consider what someone else would need to know in order to conduct this search on your behalf.
   The economic recession that suddenly occurred throughout the world in 2008 made little sense to anybody outside of economics/finance. I want to get a better understanding of how such an event can occur i.e. what features of current economics/finance allowed such a problem to happen.
3. Please specify what would constitute a relevant or non-relevant document or webpage relating to this search situation. You might want to use the format "A relevant document or website would include information about X or Y. Pages that include only information about Z are not relevant".
   A relevant document or website would include information about the recession and what principles of current economics allowed the propagation of the problem throughout the world. Pages that include only information about the period during which the recession took place are not relevant.
4. Please provide any key words or search terms you remember using or you think might be useful in searching for your topic.
   21st century recession; financial crisis 2008; financial crisis UK; Economics of recession

While the majority of respondents provided clear and detailed answers to these questions, in three cases participants described search tasks that were not clearly closed or open in nature. In these cases it was necessary to request alternative search tasks from the participants, with further guidance on the type of tasks required. The resulting data-set consisted of forty search tasks, one closed and one open from each of the twenty participants. In order to allow for comparison between primary and secondary relevance judgements, it was necessary to select eight participants at random for whose topics relevance judgements would be made by five other participants (a full explanation is provided in Section 3.4). The responses to question 1-3 were used verbatim as the structured topic description presented to participants during the lab study (see Sections 3.3 and 3.4).

3.3 Document Generation

For each participant topic, the keywords provided by participants were used as a query to conduct a web search. To present a range of different documents with potentially different degrees of relevance to the user's topic, 3 search results were sampled from each page of 10 links provided by Google, i.e. three random results were selected from page 1, then page 2, etc. until a total of 30 document results were downloaded. As far as possible, any non-HTML documents in the result list were removed from consideration during this process.

Each of these documents was then processed using the "Readability" API[1], which removed advertising and other superfluous webpage information. The aim here was to ensure that documents would be presented in a similar text and image form to each other, in order to remove issues around differing website designs. Not all of the resulting documents necessarily contained data, and so a final manual scan of the documents was carried out to remove empty documents. A final stratified sampling of 15 documents was then taken from this list, across each Google result page. This process was derived from a number of pilot tests which investigated different methods of downloading documents of different degrees of relevance. While "relevance" is not being controlled in this study, we did wish to maximise the chances of both relevant and non-relevant documents being presented to participants. Techniques to dilute search results were not found to be useful in this particular study, with the simpler assumption that documents further down the ranking were less likely to be relevant being found to provide a range of material expected by assessors. Beyond removing non-HTML documents, no attempt was made to control other document characteristics like length.

The result of this process was that for each participant topic, 15 documents were downloaded, the collection as a whole consisting of 600 unique documents. Across all topics the mean document word length was 1977 words (SD 3840). The majority of documents were less than 5000 words long (545 documents), with one document of over 50,000 words, over twice as long as any other document in the collection, belonging to topic number 8-2. A copy of all topics and documents is available for download[2].

3.4 Experimental Interface

The task description was displayed first, along with the question "How much do you know about this topic?" and the associated 7 point scale. On pressing the "view document" button the first page to be judged would be displayed. On the top right hand side of the screen a fixed dialog box asked the three questions "Is this document relevant to the topic?", "How confident are you in making this judgement?", and "Have you seen this document before?" It should be noted that participants were able, by design, to complete and submit these questions without viewing the entire document. The title of the current topic was displayed at the top of

[1] https://readability.com/developers/api

[2] http://dx.doi.org/10.15129/317def18-5702-407e-9cf4-a92ed4e6c081

the window along with a "click to view topic" button which would allow the participant to return to the full topic description

Table 1: Measures used in the study.

Measure	Description
For each document judged:	
Relevance	Binary relevance judgement (0/1).
Confidence	Degree of confidence in the relevance judgement (1 = no confidence, 7 = very confident).
Time	Time taken to make the relevance judgement.
View Topic	Number of times a user "returned" to the topic description.
For each topic:	
Knowledge	Knowledge of the topic (1 = no knowledge, 7 = expert). Recorded before judgements made.
Interest	Degree of interest in the topic (1 = not interesting at all, 7 = extremely interesting). Recorded after all judgements have been made for a topic.

At the very start of the study a single "practice" task/document was displayed, which allowed the participant to become familiar with the interface before the study topics were displayed. On commencing the study proper, the first topic description was presented. All 15 documents for each topic were then displayed in turn, with order of document presentation being randomised. After all documents for the topic had been judged the participant would then be asked "How interesting was this topic to you?" The system then moved on to the next topic, displaying the topic description followed by 15 documents. This was repeated for all 6 topics judged by each participant. All participants used the interface under the same conditions (screen/interface size and computer). The system logged mouse movement, button clicks, question responses, and other measures such as time taken for each judgement (a list of the measures used in this paper, and a subset of the full list used, is provided in Table 1). Morae was also used to record the screen of the computer.

3.5 Laboratory Protocol

The laboratory sessions were conducted in the University of Sheffield's iLab. Each participant was required to judge the relevance of fifteen documents for each of six search tasks; two being their own (one open and one closed), and four being the open and closed tasks of two other users. This meant that each participant judged the relevance of 90 documents.

Once the practice task had been completed, Morae screen recording software was started, and the participant was instructed to begin the tasks proper. No time limit was imposed for any stage of the process, and the order of tasks, and of documents within each task, was randomised for each participant. The investigator observed the session via a remote Morae connection in the iLab control room, and was able to add markers to the Morae screen recording on occasions when the participant exhibited interesting or note-worthy behaviour (for example changing their relevance judgement, making very speedy or slow judgements, or assigning a low confidence value to their judgement).

On completion of all six tasks, the investigator returned to the lab and loaded the Morae screen recording onto the participant's PC. The participant was then asked to watch back their session and describe their behaviour and the rationale behind their relevance judgements. Due to time constraints it proved impractical for the participant to watch the whole of their session. Instead particular attention was paid to the first documents for each task, and other documents that the investigator had marked as noteworthy. Participants were also asked explicitly for their perspective on the differences between completing their own and other participants' tasks, and on the perceived effects of topic knowledge and interest. Attention was also paid to their interpretation of the confidence scale. All replay and interview sessions were recorded, and the audio recordings transcribed. The transcriptions were then subjected to Qualitative Content Analysis.

3.6 Demographics

The experiment had 20 participants, who were predominantly staff and students at the University of Sheffield. The participants had an average age of 27.9 (std. dev. = 8.17), the youngest participant was 19 and the oldest 54. 9 of the participants were male and 11 female. 12 of the participants were native English speakers; the other native languages were Indonesian, Japanese, Hindi, Chinese, Arabic and Italian. Of the non-native speakers, 5 rated their English as fluent and 3 at an intermediate level. In terms of search experience, 14 participants reported that they had high search experience, with the remaining 6 reporting medium experience. In the experiment 1800 relevance assessments were made in total. Of those 1200 were secondary assessments and 600 primary. 240 documents received more than 1 assessment, and each of those documents received 1 primary and 5 secondary assessments, giving a total of 1440 assessments.

4. RESULTS

4.1 How does relevance assessment behaviour differ between primary and secondary assessors?

4.1.1 Quantitative Results

Table 2: Summary of results for primary and secondary judgements. Significant differences in bold

	Primary		Secondary	
	Median	Mean (SD)	Median	Mean (SD)
Relevant	0	0.448 (0.498)	0	0.447 (0.497)
Time (millisecs)	23687	34630 (33982)	26496	38727 (40263)
Milliseconds per word	**34.045**	**53.482 (96.137)**	**38.336**	**68.349 (123.152)**
View Topic	**0**	**0.04 (0.212)**	**0**	**0.135 (0.388)**

Table 2 presents a summary of some assessor behaviour statistics, showing the number of documents marked relevant, absolute time to make a judgement, time to make a judgement scaled by the word length of the document, and number of times the 'view topic' button was pressed. As the data was not normally distributed Wilcoxon Rank Sum tests were used to compare between conditions. Significant differences were found for scaled time ($p = 0.01$, $W = 386670$) and number of view topic button presses ($p < 0.001$, $W = 390312$). We show absolute times in addition to scaled time since previous work has shown that document length can affect effort (e.g. [18]). It was found that assessors pressed the 'view topic' button more often on secondary topics, although the relatively rarity of this action results in very small per-session

numbers. In total, across the whole data set, there were 162 button presses for secondary topics (out of 1200 document sessions) versus only 24 button presses for primary topics (out of 600).

4.1.2 Qualitative Results

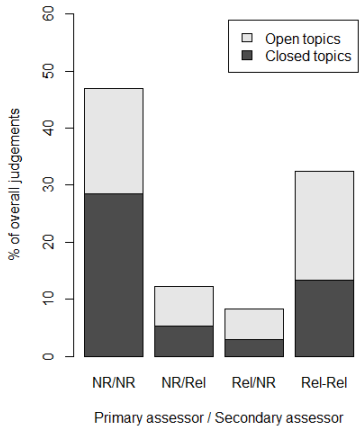
Several participants stated during the post-session interview that they felt other people's tasks took longer to complete. A number of different explanations for this were given. A common theme related to difficulties understanding the scope and details of the topic instructions, which led to extra time considering the relevance of individual documents as well as time taken to review the topic description. Lack of familiarity with the subject of the topic was also mentioned as a factor, leading to difficulties unravelling complex vocabulary and terminology, and determining the most appropriate keywords for which to scan documents. In practice this often meant that documents had to be read in closer detail, or scanned several times for different terms:

"I just had to like put in a lot of effort. I needed to try and work out what words to look for, or I needed to go through the entire thing to be sure if it's relevant."

It should also be noted that a number of participants also stated that they felt their speed increased over the course of the experiment as they became more comfortable with the requirements of the study.

4.2 To what extent do secondary assessments agree with primary judgements?

4.2.1 Quantitative Results



	NR/NR	NR/Rel	Rel/NR	Rel-Rel
Closed topics	57%	11%	6%	27%
Open topics	37%	14%	11%	38%
All topics	47%	12%	8%	32%

Figure 1: Overall agreement between primary and secondary assessors.

Within the data set there are 16 topics (8 closed and 8 open) which have a primary judgement and a total of 5 other secondary judgements by other assessors. Figure 1 shows the overall percentage agreement between primary and secondary assessors, plus the split between the open and closed tasks. Over all topics, the overall agreement was 79%, rising to 84% for closed topics and falling to 75% for open topics. This agreement is somewhat higher than that reported by Alonso and Mizzaro [3] when using crowdsourcing.

The first column of Table 3 gives the Fleiss' kappa between all assessors (column 1) and between secondary assessors only (column 2). Both follow a similar pattern, with an overall kappa value of 0.545 for all topics, indicating fair to good agreement. Again, this was higher than the study of Alonso and Mizzaro [3]. Overlap was also calculated (the intersection between secondary and primary assessors divided by the union of the relevance assessments), following Voorhees [20] with results presented in the final column of Table 3. The overall overlap was 0.61, which is a higher agreement than that reported by Voorhees [20].

Table 3: Fleiss' kappa for all assessors (primary + secondary), only secondary assessors, plus the overlap between primary and secondary.

	Kappa (all)	Kappa (secondary)	Overlap
Closed topics	0.589	0.565	0.62
Open topics	0.483	0.469	0.61
All topics	0.545	0.526	0.61

There was considerable variation in the number of documents considered relevant for each topic (Figure 2). Topics such as 5-1 and 9-1 contained many "non-relevant" documents judged by the primary assessor, as indicated by the dark bars. Other topics contained far more relevant documents, e.g. topic 16-2 or 6-2. These differences are at least partly due to the quality of the search engine results presented to assessors, and potentially suggests that the data set contains a range of different topic difficulties.

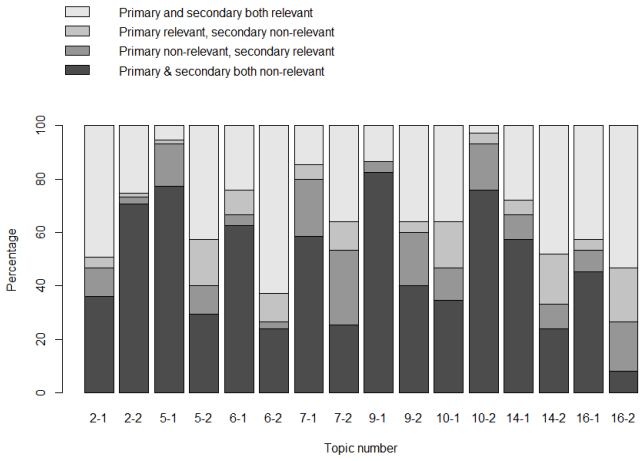


Figure 2: Percentage agreement across topics. Closed topics are 2-1, 5-1, etc., open topics are 2-2, 5-2, etc. The first number is the ID of the participant.

Following Alonso and Mizzaro [3], we also show the group agreement in Figure 3. This is calculated by subtracting the mean relevance judgement of all 5 secondary assessors from the primary relevance judgement (which can be 0 or 1). E.g. if the primary assesses a document as being non-relevant and the mean of the secondary assessors is 0.6 (3 relevant and 2 non-relevant, or 3/5), then the "error" is -0.6. As can be seen in Figure 4, in many cases all 5 secondary assessors exactly agree with the primary (in 49% of all cases). In 23% of cases (ranging from -0.2 to +0.2 in Figure 3, only a single assessor disagrees with the primary, while in 7% of cases two assessors disagree with the primary. If we were to use a "majority vote" to determine relevance from the 5 secondary

judgements, in 79% of all cases the majority would match the primary assessment. For comparison with Alonso and Mizzaro [3], the solid red box and dotted red box represent where one assessor (solid box) or two assessors (dashed box) have disagreed with the other assessors (and the primary assessor).

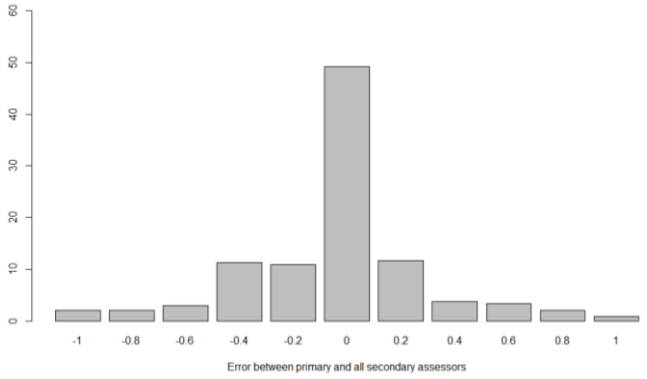


Figure 3: Difference between all secondary assessors and the primary.

4.2.2 Qualitative Results

The post-session interviews provided a rich source of data relating to perceived differences in the primary and secondary judgement processes. Perhaps most interesting were participants' comments about their interpretation of other people's topics, and in particular determining the appropriate criteria for a relevant document. As one participant put it:

"Sometimes I found it really hard to work out what they wanted. It would seem clear at first but then you'd see a document and get confused about whether was on exactly what they were looking for"

For many participants this became a question of whether to take a literal or broad interpretation of the topic description. Several participants described encountering documents that they felt might be useful and relevant to the general topic, but that did not meet the precise criteria laid out in the description. The following extract is typical of several exchanges during the post-session interview and replay assessments:

Interviewer: I think I noticed that you judged this document as relevant.
Participant: Yes, I just felt, well it's got quite a lot of useful information in it.
Interviewer: But if we look at what the actual task was…
Participant: I don't think it does actually have the exact information I was looking for, does it?
Interviewer: Yes, right, you were looking for the release dates of these two films.
Participant: Yes, but I thought, "Oh that's interesting." I think this would be useful for the person doing the search even if it didn't have everything they wanted.

This had interesting implications when primary assessors encountered such situations. They described consciously deciding whether to limit themselves to the confines of the topic description, or taking a more "real-life" approach: *I was almost thinking, "Am I coming at it from being me and knowing what I was looking for or am I coming at it being a participant going off the description of what it was?"*

Primary assessors frequently admitted some prior experience of evaluating documents on a similar topic to the one in the study. This led to situations where the novelty of a document became a factor in the relevance assessment, something which was clearly not applicable to secondary judgements.

In almost all cases, secondary judgements were deemed to be more difficult than primary ones. Aside from issues relating to topic scope, and the keyword identification issues described above, the characteristics of the documents themselves were sometimes a cause of difficulties. Some participants stated they were unsure whether the reliability or source of a document should be considered. These assessors spoke of doubts about whether documents that were recognisable as blogs or opinion pieces should be considered relevant even if they appeared to be topical:

"I think it does affect how you look at the document, how much weight you put to it and the confidence you would take from it. So ultimately whether it's relevant or not."

It should also be noted that participants described substantial differences between open and closed tasks. Closed tasks were almost universally perceived as easier, with the Open tasks were viewed as more subjective, and as such were more prone to the secondary judgement issues described above:

"Open tasks were just more open to interpretation, just what you wanted as a person to find out. Which was hard when it's not what you specifically want to find out."

4.3 How do contextual factors such as interest in the topic and knowledge of the topic affect relevance judgements?

4.3.1 Quantitative Results

Table 4: Differences between primary and secondary assessors for topic knowledge and interest.

	Primary		Secondary	
	Median	Mean (SD)	Median	Mean (SD)
Knowledge	6	5.25 (1.375)	1	2.062 (1.461)
Interest	6	5.475 (1.55)	4	3.513 (1.636)

Summary statistics for topic knowledge and interest in the topic for all assessors are shown in Table 4. Wilcoxon rank sum tests found that there were significant differences between primary and secondary assessors for both knowledge ($p < 0.001$, $W = 60522$) and interest ($p < 0.001$, $W = 136688$). As can be seen in Table 4 primary assessors were significantly more knowledgeable and interested in their own topics. Figure 4 shows the relationship between knowledge and interest as reported by secondary assessors versus the Cohen Kappa agreement between these secondary judgements and the primary judgements. Kruskal-Wallis tests were used to investigate significant relationships, with none being found, i.e. neither greater interest nor greater knowledge in the topic resulted in greater agreement between the secondary and primary assessors. This analysis was also repeated for open and closed tasks, again with no significant results being found. Histograms showing the distribution of secondary topic knowledge and interest are also shown in Figure 4. Few secondary assessors used the top of the knowledge scale, it being highly skewed right with both a median and mode of 1. For topic interest the distribution is almost constant through the first six levels with the exception of the top 7 rating which not selected by any secondary assessor. This would suggest secondary assessors were reluctant to indicate that they were "experts/knowledgeable" in a topic, but were much more

likely to indicate that they were "interested" in the same topic, but not "extremely interested".

4.3.2 Qualitative Results

Two key themes emerged during the post-session interviews. First, that participants were generally much more interested in their own tasks, which meant they were happy to spend longer reading documents where necessary, and were less likely to feel frustrated during the judgement process. Two participants described instances of encountering documents that would be of use in real-life:

"I was bored before I started this task. I've just got no interest in it so it was hard to care whether a document was relevant or not."

"So I just took a while and I was like, "Okay. Let me note that. I will go home and look for them." (Laughter)

Second, that participants were clear that they were likely to have greater topic knowledge for their own tasks. Participants frequently saw this factor as the main cause of it being easier to identify keywords for their own tasks, and were less likely to be troubled by specialist vocabulary present in some documents*: "It's easier for me to absorb because I know more about environmental issues than I do about international finance and banking structures."*

4.4 Does the length of the topic description affect secondary relevance judgements?

Table 5 shows the word length of the different components of the topics, the description, situation, and criteria parts, as well as the mean topic length in words. On average topic size was roughly in line with TREC-5 (mean 82.7 words per topic) and TREC-6 (mean 88.4 words per topic [19]). On average open topics were slightly longer than closed.

Table 5: Mean (SD) length of topic in words, spit be different section of topic and open/closed

	Description	Situation	Criteria	All
Closed	9.8 (4.3)	41.6 (8.8)	28.4 (11.5)	79.8 (13.1)
Open	11.9 (5.3)	43.1 (16.8)	36.9 (11.5)	91.9 (18.5)
All	10.8 (5.0)	42.4 (13.3)	32.6 (12.2)	85.8 (17.1)

Al-Harbi and Smucker [2] suggested that there may be a relationship between topic length and number of documents judged relevant by assessors, arguing that shorter descriptions resulted in broader interpretations of relevance criteria, and therefore a higher number of documents judged relevant. Figure 5 illustrates this relationship (for all assessors, and also split between primary and secondary assessments). Spearman's rank correlation coefficients were used to test these relationships and no significant correlations were found, i.e. the length of the topic description did not appear to affect the number of documents marked relevant by assessors. A similar analysis for carried out for only open and closed topics, and again no relationship was found.

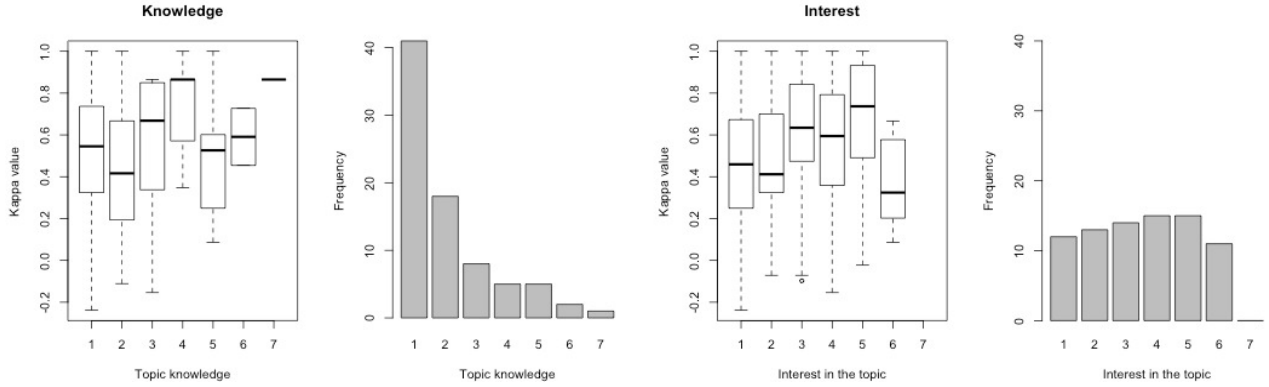


Figure 4: Secondary assessor agreement with primary by topic knowledge and topic interest. The distribution of knowledge and interest for all secondary judgements are also shown.

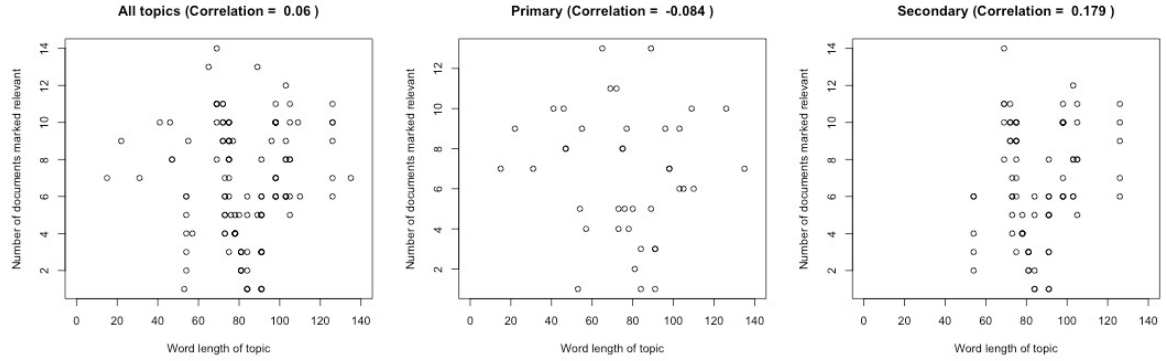


Figure 5: Number of documents marked relevant versus the length of the topic description (all, primary judgements only, and secondary judgements only, from left to right)

4.5 How does confidence in judgements differ between primary and secondary assessors?

4.5.1 Quantitative Results

Table 6: Confidence values for assessors who have judged primary and secondary documents, Mean (SD). Significant differences in bold.

User	Primary	Secondary	P value	W
2	6.867 (0.434)	6.683 (0.624)	0.132	1020.5
5	**6.9 (0.403)**	**5.667 (1.515)**	**< 0.001**	**1376**
6	**6.667 (0.606)**	**5.333 (1.515)**	**< 0.001**	**1371**
7	6.267 (1.143)	6 (1.414)	0.485	974
9	**6.833 (0.747)**	**6.5 (0.893)**	**0.022**	**1091.5**
10	**6.133 (1.306)**	**5.283 (1.151)**	**< 0.001**	**1315.5**
14	6.933 (0.365)	6.75 (0.571)	0.057	1031.5
16	5.9 (1.125)	5.9 (1.298)	0.698	856.5
All	***6.302 (1.036)***	***5.712 (1.4)***	***< 0.001***	***273891***

Table 7: Confidence for primary vs secondary assessors for the 16 topics with secondary assessments, mean (SD). Significant differences in bold, "closed" topics are 2-1, 5-1, etc., "open" topics numbered 2-2, 5-2, etc.

Topic	Primary	Secondary	p-value	W
2-1	6.8 (0.561)	6.373 (1.148)	0.201	653.5
2-2	**6.933 (0.258)**	**6.533 (0.741)**	**0.042**	**710.5**
5-1	**6.933 (0.258)**	**5.467 (1.711)**	**< 0.001**	**878**
5-2	**6.867 (0.516)**	**5.827 (1.288)**	**< 0.001**	**845.5**
6-1	**6.8 (0.561)**	**5.987 (1.257)**	**0.007**	**792.5**
6-2	6.533 (0.64)	5.96 (1.38)	0.244	661.5
7-1	**6.733 (0.594)**	**5.28 (1.122)**	**< 0.001**	**964.5**
7-2	**5.8 (1.373)**	**4.973 (1.174)**	**0.012**	**787**
9-1	6.933 (0.258)	6.88 (0.366)	0.639	585.5
9-2	**6.733 (1.033)**	**5.613 (1.46)**	**<0.001**	**866**
10-1	6.133 (0.99)	5.4 (1.533)	0.114	703.5
10-2	**6.133 (1.598)**	**5.587 (1.347)**	**0.043**	**742.5**
14-1	**6.867 (0.516)**	**6.107 (1.247)**	**0.007**	**783**
14-2	**7 (0)**	**4.613 (1.895)**	**< 0.001**	**1012.5**
16-1	6 (1.134)	5.587 (1.14)	0.167	685.5
16-2	**5.8 (1.146)**	**5.2 (1.027)**	**0.027**	**759**

Overall it was found that primary assessors were more confident in their judgements when compared to secondary assessors (Wilcoxon rank sum test was significant $W = 273891$, $p < 0.001$). Table 6 shows the overall mean and SD confidence values for all users (final row), and also for the assessors who have judged both primary and secondary documents. As can be seen, for four assessors confidence on primary assessments was significantly higher when compared to that assessor's confidence on the secondary assessments. For the four other assessors, however, there were no differences in confidence between primary and secondary assessments. While confidence was generally high across all users, this also varied (e.g. user 16 in Table 6).

Looking at confidence by topic, Table 7 shows the confidence split by the 16 topics for which there are primary and secondary assessments. Wilcoxon rank sum tests were used to compare primary and secondary confidence values, with significant differences being found for 11 topics (p-values and W test statistics are shown in Table 7). It should be noted that while Table 6 compares a single assessor's confidence on primary topics vs. the same assessor's confidence on secondary topics, Table 7 compares the confidence of the primary assessor against the other five secondary assessors.

Across topics confidence is generally high, but there are some striking differences between primary and secondary assessments, such as Topic 14-2. Topic 14-2 was an open topic, with description "What led to the recession that began in 2008?" While the primary assessor was obviously confident in his/her judgements, the same could not be said of the secondary assessors. For other topics confidence was almost equal between primary and secondary assessors, e.g. topic 9-1, a closed topic with description "What year was the original Vienna State Opera House completed?" In this case both primary and secondary assessors indicated that they were uniformly confidence in their relevance judgements. Overall, it was found that confidence on closed and open tasks did vary significantly (Wilcoxon rank sum test $p < 0.001$, $W = 453304$), although as can be seen in Table 7 this also varied by topic. For the 16 topics in Table 7 in 5 cases there was no significant difference between primary and secondary assessors, four out of the five being closed topics. For the other 11 topics where differences were found 7 were open topics and 4 closed.

4.5.2 Qualitative Results

A number of interesting findings emerged from the qualitative data. It was notable that in many cases, participants struggled to explain both how they interpreted the confidence scale, and the factors that influenced their confidence judgement:

"It was very difficult. I think it was just subjective, I think it was just depending on what I felt."

Those participants who were able to articulate their assessment of confidence described a range of factors influencing their confidence score, including the speed with which they were able to make their relevance judgement, the reliability of the source, how well they felt they understood the document, and how clearly they understood the topic description:

"How easy it was to relate to what the situation was, what they were searching for. I think I was confident in most of them and close to very confident in most of all my judgements."

"If I find something that is relevant but I cannot totally understand the document, I will choose less confident."

"It was partly down to just how reliable I thought the document was."

Perhaps most striking was the number of participants who described using the confidence scale as a proxy for graded relevance:

"In many ways I was using that confidence scale as more of a precise relevance scale. It was how relevant I thought it was."

In total over half of participants equated the confidence value with a measure of the document's relevance.

5. DISCUSSION

Before addressing the question of agreement levels between primary and secondary assessors for real-life search tasks, it is instructive to review results of this study relating to the judgement process itself. Looking first at speed of judgement, we note that when scaled by document length, primary assessors were found to be significantly quicker in making their judgements. Substantial differences in speed were also observed between open and closed search-tasks. Participants were quicker to judge the relevance of documents relating to closed tasks, and differences between the two types of task were also mentioned by participants in the post-task interviews. Judgements for open tasks were seen as more difficult to make, and the judgement process itself was considered more taxing. This was in part due to the additional factors perceived as influencing relevance for open tasks such as the reliability of the document. Given Yilmaz et al.'s [26] findings showing the relationship between effort, relevance, and utility we observe that using open search tasks for judging relevance within test collections may result in relevance judgements based on factors beyond topicality.

Results of this study also confirm that for real-life search tasks, knowledge of and interest in the topic are greater for primary assessors than secondary. However, as shown in Section 4.3, no results were found which suggested that an increase in the interest or knowledge of a secondary assessor would increase the chance of the secondary agreeing with a primary. While this may partially be a consequence of self-reporting scales (we note for example that even primary assessors rarely ranked their knowledge of a topic highly), our results do suggest that secondary assessors are generally well able to make topical relevance judgements even while professedly unsure of the full scope, context or background to a topic.

The post-session interviews revealed that the form and complexity of the topic description was a key factor affecting secondary assessment. The topic descriptions which were gathered in this study turned out to be roughly the same length as many TREC topics (Table 5). It has been suggested that topic length may be related to number of documents marked relevant, but we could find no evidence of this in our data set. However given the qualitative results in Section 4.2.2, taking a simple word count for a topic may not be a good representation of the complexity or difficulty of that topic to an assessor, a view supported by Bell and Ruthven [5]. A key theme to emerge from the interviews was the difficulty secondary assessors had in interpreting the context and scope of other participants' open tasks from the task description text. It seems likely that many of the disagreements in relevance judgement were a consequence of how secondary assessors chose to construe the task description. This is supported by the data showing secondary assessors were significantly more likely than primary assessors to review the topic description while undertaking relevance judgement, and suggests that the form and content of task descriptions can play an important role in minimising the interpretative challenges faced by secondary assessors. We suggest that further research investigating the precise effect of variations in task description structure and content could provide valuable insight into optimising task descriptions for secondary relevance assessments. A significant difference was found between the confidence of relevance judgements between primary and secondary (see Figure 7), with primary assessors being generally more confident in their judgements. However the difference was not large, and assessors in general were found to be confident in their judgements. This is a somewhat surprising finding given that the use of a binary rather than graded relevance scale forced assessors to resolve doubts about a borderline document one way or the other. One explanation for this can be found in the interview data, which suggests that the confidence scale used was problematic: different assessors used the scale in different ways, and some found it extremely difficult to articulate both the factors influencing the certainty of their judgement, and the way in which the confidence scale was interpreted. Although not the original focus of this research, we conclude that for many assessors, understanding and measuring judgement certainty is problematic, particularly if required to use a Likert-type scale. We suggest that further work investigating more effective means of soliciting a measure of judgement confidence might be of considerable value.

We find then that when judging the relevance of documents for real-life search tasks, secondary assessors are less knowledgeable, find the process slower and more demanding, perceive the topic as less interesting, and are less confident in their judgements. They also face substantial problems interpreting the scope of the topic, and determining the criteria for relevance. Yet despite these apparently confounding factors, agreement levels between primary and secondary assessors were found to be high. Comparing our results with studies investigating non-real-life search tasks, we observed a greater overall level of agreement with the primary (79%) than [3] (68%), and a similar level of majority agreement. We also found a higher level of judgement overlap (.61) than [20] (.30). There is little doubt that this is at least in part due to the nature of the topics and documents under consideration, which of course differed from standard TREC evaluations used in [3; 23], and the characteristics of the assessors and assessment environment (we note in particular here that [3] were utilising crowd-workers). Nonetheless, we have shown that secondary assessors produce high levels of agreement with the creators of real-life search tasks. Put another way, we find that the judgements of assessors for whom the topic does not represent a real-life information need are generally the same as those for whom it does. Given the many practical difficulties of obtaining test-collection judgements from real-life topic owners, it is reassuring to conclude that using synthetic search tasks is unlikely to affect judgement quality, and by extension the accuracy of laboratory evaluations.

6. CONCLUSIONS AND FUTURE WORK

The concept of relevance continues to be of importance to information retrieval and information science research. Much research in this area has involved the use of TREC topics. Unfortunately, as pointed out elsewhere [2], these topics are dated. One of the aims of this work has been to revisit relevance assessment using up to date "real-life" topics gathered from staff and students at a major university. From the data collected, it is possible to gain a greater understanding of such real-life assessments, and enables us to compare our results to the large volume of previous work which has used TREC.

While behavioural differences were found between primary and secondary assessors (e.g. time to judge when scaled by document length) agreement between primary and secondary assessors was generally high. Self-reported contextual factors (topic interest and knowledge) did not appear to affect assessor agreement. This was despite secondary assessors generally assessing themselves as being less knowledgeable and less interested in the topics, and qualitative results suggesting that assessors found the relevance assessment task difficult. In attempting to interpret these results it is important to acknowledge some limitations of this study. In particular we note that while primary assessors were assessing the relevance of documents to their real-life search tasks, the judgement process itself was essentially artificial, since it occurred

under laboratory conditions and using a constructed document set. We therefore emphasise that these results are most usefully interpreted within the context of the standard relevance assessment process for IR system test collection development. In this sense, our results support the notion that the use of synthetic topics for relevance assessment, as typified by TREC, result in judgement sets of no lower quality than those for real-life topics.

One final result of this work worthy of discussion concerns the instruments which we use to gather more information about the relevance judgement process itself. While measuring the confidence of an assessor in a relevance judgement is an intuitively attractive proposition, results here suggest that in practice its use can be problematic. Assessors were found to interpret a simple seven point scale in very different ways, including as a proxy for graded relevance. We believe that investigating novel ways of measuring factors such as confidence which do not themselves become proxies for relevance is a subject for future work.

7. ACKNOWLEDGEMENTS

This work was funded by a UK Arts and Humanities Research Council grant to the second and third authors (grant AH/L010364/1).

8. REFERENCES

[1] Agosti, M., Fuhr, N., Toms, E., and Vakkari, P. 2014. Evaluation methodologies in information retrieval (Dagstuhl Seminar 13441). *SIGIR Forum* 48, 1, 36-41.

[2] Al-Harbi, A.L. and Smucker, M.D. 2014. A qualitative exploration of secondary assessor relevance judging behavior. In *Proceedings of the 5th Information Interaction in Context Symposium* (IIiX '14). ACM, New York, NY, USA, 195-204.

[3] Alonso, O. and Mizzaro, S. 2012. Using crowdsourcing for TREC relevance assessment. *Inform. Process. Manag.* 48, 6, 1053-1066.

[4] Bailey, P., Craswell, N., Soboroff, I., Thomas, P., De Vries, A.P., and Yilmaz, E. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '08). ACM, New York, NY, USA, 667-674.

[5] Bell, D.J. and Ruthven, I. 2004. Searcher's assessments of task complexity for web searching. In *Proceedings of the 26th Annual International European Conference on Information Retrieval* (ECIR 2004), Springer-Verlag, Berlin, Germany, 57-71.

[6] Borlund, P. 2003. The IIR evaluation model: a framework for evaluation of interactive information retrieval systems. *Information Research 8*, 3.

[7] Chouldechova, A. and Mease, D. 2013. Differences in search engine evaluations between query owners and non-owners. In *Proceedings of the sixth ACM international conference on Web search and data mining* (WSDM '13). ACM, New York, NY, USA, 103-112.

[8] Clough, P., Sanderson, M., Tang, J., Gollins, T., and Warner, A. 2013. Examining the limits of crowdsourcing for relevance assessment. *Internet Computing, IEEE.* 17, 4, 32-38.

[9] Efthimiadis, E.N., and Hotchkiss, M.A. 2008. Legal discovery: Does domain expertise matter? *P. AM. SOC. INFORM. SCI.* 45, 1, 1-2.

[10] Jones, K.S. 1998. Further reflections on TREC. *Inform. Process. Manag.* 36, 1, 37-85.

[11] Kelly, D. 2009. Methods for evaluating interactive information retrieval systems with users. *Foundations and Trends in Information Retrieval 3*, 1—2, 1-224.

[12] Kinney, K.A., Huffman, S.B., and Zhai, J. 2008. How evaluator domain expertise affects search result relevance judgments. In *Proceedings of the 17th ACM conference on Information and knowledge management* (CIKM '08). ACM, New York, NY, USA, 591-598.

[13] Marchionini, G. 1989. Information-seeking strategies of novices using a full-text electronic encyclopedia. *J. Am. Soc. Inf. Sci. Tec.* 40, 1, 54-66.

[14] Ruthven, I., Baillie, M., and Elsweiler, D. 2007. The relative effects of knowledge, interest and confidence in assessing relevance. *J. Doc.* 63, 4, 482-504.

[15] Saracevic, T. 1995. Evaluation of evaluation in information retrieval. In *Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '95), ACM, New York, NY, USA, 138-146.

[16] Saracevic, T. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part III: Behavior and effects of relevance. *J. Am. Soc. Inf. Sci. Tec.* 58, 13, 2126-2144.

[17] Smucker, M.D., and Jethani, C.P. 2011. Measuring assessor accuracy: a comparison of nist assessors and user study participants. In *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval* (SIGIR '11). ACM, New York, NY, USA, 1231-1232.

[18] Villa, R. and Halvey, M. 2013. Is relevance hard work?: evaluating the effort of making relevant assessments. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 765-768.

[19] Voorhees, E., and Harman, D. 2000. Overview of the Seventh Text Retrieval Conference. In *Proceedings of the Seventh Text REtrieval Conference (*TREC-7*)*, NIST Special Publication, 1-24.

[20] Voorhees, E. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inform. Process. Manag.* 36, 5, 697-716.

[21] Webber, W., and Pickens, J. 2013. Assessor disagreement and text classifier accuracy. In *Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '13). ACM, New York, NY, USA, 929-932.

[22] Webber, W., Toth, B., and Desamito, M. 2012. Effect of written instructions on assessor agreement. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval* (SIGIR '12). ACM, New York, NY, USA, 1053-1054. |

[23] Yilmaz, E., Verma, M., Craswell, N., Radlinski, F., and Bailey, P. 2014. Relevance and effort: an analysis of document utility. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management* (CIKM '14). ACM, New York, NY, USA, 91-100.