A Qualitative Exploration of Secondary Assessor Relevance Judging Behavior

Aiman L. Al-Harbi
School of Computer Science
University of Waterloo, Canada
a2alharb@uwaterloo.ca

Mark D. Smucker
Department of Management Sciences
University of Waterloo, Canada
mark.smucker@uwaterloo.ca

ABSTRACT
Secondary assessors frequently differ in their relevance judgments. Primary assessors are those that originate a search topic and whose judgments truly reflect the assessor’s relevance criteria. Secondary assessors do not originate the search and must instead attempt to make relevance judgments based on a description of what is and is not relevant. Secondary assessors may be hired to help in the construction of test collections. Currently our knowledge about secondary assessors is largely limited to quantitative measurements of the differences between judgments produced by secondary and primary assessors. In order to better understand the behavior of secondary assessors, we conducted a think-aloud study of secondary assessing behavior. We asked secondary assessors to think-aloud their thoughts as they judged documents. The think-aloud method gives us insight into how relevance decisions are made. We found that assessors are not always certain in their judgments. In the extreme, secondary assessors are forced to make guesses concerning the relevance of documents. We present many reasons and examples of why secondary assessors produce differing relevance judgments. These differences result from the interactions between the search topic, the secondary assessor, the document being judged, and can even apparently be caused by a primary assessor’s error in judging relevance. To improve the quality of secondary assessor judgments, we recommend that relevance assessing systems allow for the collection of assessor’s certainty and provide a means to help assessors efficiently express their judgment rationale.

Categories and Subject Descriptors
H.3.4 [Information Storage and Retrieval]: Systems and Software Performance evaluation: Effectiveness

Keywords
Search, Evaluation, Secondary Assessors

1. INTRODUCTION
In information retrieval (IR), evaluating the effectiveness of an IR ranking algorithm is commonly carried out by using test collections and associated effectiveness measures. A typical test collection consists of three parts: (1) documents, which are the objects to be judged; (2) search topics, which represent the information need; and (3) relevance assessments, which denote the relevance of each document in the test collection to a particular search topic [21]. Relevance judgments are made either by primary assessors or secondary assessors. Primary assessors are the ones who create the search topic, while secondary assessors are simply hired to make relevance judgments and are not, therefore, the search topics’ originators. In many cases secondary assessors lack the expertise or knowledge of primary workers and are thus more affordable. Another motivation for use of secondary assessors is that judging can be completed faster if multiple assessors work on each search topic.

Different assessors produce different relevance judgments. Much of the existing research on secondary assessors and the differences in judgments produced by them has focused on the examination of specific causes of these differences. In this paper, we provide a broader view of the the relevance judging behavior of secondary assessors.

To gain insight into secondary assessor behavior, we conducted a qualitative study where we asked study participants to think aloud their thoughts as they made relevance judgments. Since the aim is to investigate secondary assessors’ judging behavior in a broader view, we did not limit ourselves in this study to certain factors. We selected documents and search topics from the TREC 2005 Robust track [20]. In selecting documents, we chose a mixture of documents for which we already knew from prior studies that secondary assessors either consistently agree on the relevance of the document or lack consensus on the document’s relevance. In line with the previous studies, the majority of differences in relevance judgments came from the low consensus documents. In addition to the judgments, we observed the participants first hand and audio recorded their think-aloud session as well as the video of the computer. The recordings were transcribed, and based on the transcriptions and the other observations, we discovered that:

- Secondary assessors make many judgments with less than full certainty. While we may collect binary or graded judgments, the reality is that relevance judgments are not always absolute in the minds of secondary assessors. Given that secondary assessors are making uncertain judgments, we find it hard to de-
scribe variations in relevance judgments as purely disagreements concerning the meaning of relevance.

- Differences between primary and secondary assessors can be divided into four categories: difficulty in applying the search topic, difficulty in processing the document, secondary assessor factors, and true error in the primary assessor’s judgment.

- The length of a search topic’s description may influence the nature of judgment differences. We found preliminary evidence that a short topic description appears to lead to a liberal relevance criteria where more documents appear to be relevant to the secondary assessor; a long topic description appears to lead to more conservative behavior where fewer documents appear to be relevant.

In the next section, we review related work before describing our experiment and its results in more detail.

2. RELATED WORK

Disagreeing about relevance judgments by assessors is not a recent concern in IR research. It has been observed and studied since the 1960s and 1970s [5, 11]. Lesk and Salton [11] and Cleverdon [5] studied the impact of different relevance judgment sets on the performance of retrieval systems. They found that the stability of retrieval systems’ performance was not impacted even though different relevance judgment sets were used. Likewise, Voorhees [19] studied the impact of different relevance judgment sets on the performance of retrieval systems but this time with larger test collections. Ultimately, she reached similar results to that of Lesk and Salton and Cleverdon. Harter [9] conducted a study where he examined a wide number of past studies that discuss the impact of variations in relevance judgments on measures of retrieval effectiveness. He concludes by emphasizing on the importance of enhancing the evaluation approaches that consider carefully the variations in relevance judgments, human factors, and individual differences.

Bailey et al. [1] concluded that assessors who are experts have a tendency to make more accurate relevance judgments than non-expert assessors. Wang [22] and together with Sofergel [23] also examined if background has an impact on assessors’ relevance judgment agreements. They found that even though the two groups were from different disciplines: law and library information studies (LIS), both produced almost identical relevance judgments. However, the LIS group judged the non-relevant documents a bit less accurately than the law group. Efthimiadis and Hotchkiss [7] asked assessors to judge the relevance of documents for search topics that are designed for the Legal Track Interactive Task Challenge. They found that assessors who did not have legal expertise were much better at assessing the relevancy of documents than those having legal expertise. Grossman and Cormack [8] found that most of the disagreements in relevance assessments are due to human error. Also, Villa and Halvey [18] found that assessors tend to pay much greater effort as the document increases in size. Moreover, documents that are either highly relevant or not relevant needed less effort to be assessed. The level of accuracy, while not impacted by the length of judged documents, was impacted by the degree of relevance of a document. Scholer et al. [13] showed that the first documents that an assessor encounters during the relevance judging process will affect his or her relevance threshold. In another work, Scholer et al. [14] pointed out how the same assessor might produce different relevance assessments for the same document at different points of time during the relevance judging process. In an effort to predict the disagreement in assessors’ relevance judgments, Chandar et al. [4] designed a study to predict assessor disagreement. In this study, two of their hypotheses were found to be true: (1) “longer documents will provoke more disagreement”, and (2) “less coherent documents will provoke more disagreement”; however, and surprisingly, a third hypothesis found that “documents that are more difficult to read will provoke higher levels of assessors disagreement” was found to be untrue. In fact, they found the opposite to be true: “documents that are easier to read are the ones that provoke more assessor disagreement”.

Webber et al. [24] found that the probability of assessor disagreement varies based on the rank of the retrieved documents. Highly ranked relevant documents and low ranked non-relevant documents have the least amount of disagreement while highly ranked non-relevant documents and low ranked relevant documents produce the most disagreement.

The study of Ruthven and Elsweiler [12] is an example of studies that focus on studying the impact of certain factors on the relevance judging process. They conducted a study to investigate the impact of three factors: (1) the assessor’s level of knowledge of a search topic, (2) his interest in the search topic, and (3) his confidence level when assessing the relevancy of documents. Based on their results, all of these factors were found to have an impact on assessors’ relevance judgments.

In summary, existing research has shown that judgment differences may result from differences in assessor expertise, the nature of the documents being judged, the order of documents judged, and assessor error. Our work confirms several of these causes of assessor judgment differences while providing new insights into assessor behavior.

3. METHODS

3.1 Think-Aloud Protocol

We conducted a think-aloud study. In think-aloud studies, users are asked to speak aloud thoughts while they work [17]. This method gives insight into what is going on in participants’ minds as they complete a task. The research facilitator does not intervene while the assessors work on the assigned tasks. The research facilitator’s only task is to encourage participants to continue speaking if there is a lengthy period of silence.

3.2 Study Design

We created an experiment that consisted of two parts: The Tutorial Part, which helped train the participants in how to judge the relevancy of documents; and The Task Part, which we used to collect real data. In this subsection, we describe the methods and materials used in building the experiment and collecting the data.

3.2.1 Selection of Search Topics and Documents

In total, five search topics were used in this study. We used four for the main search tasks in the Task Part, and one for the tutorial in the Tutorial Part. Each was taken from the TREC 2005 Robust Track [20]. The topics that
we used in the Task Part of the study, and their titles, are shown in Table 1. We used topic 427 (UV Damage, Eyes) for the tutorial.

When choosing documents for the study, we relied on data collected in previous studies [10, 15, 16]. In these studies, documents were judged by many participants; therefore, we have a good level of knowledge about their consensus levels. We refer to a consensus level here as the general agreement upon the relevancy of a document to a given search topic. For instance, if a document was judged relevant or not relevant by 90% of assessors, then the majority of assessors agreed on its relevancy either relevant or not relevant; therefore, it is a high consensus document. In contrast, if just 50% of assessors judged it as relevant or not relevant, then we do not have a general agreement on its relevancy and as a result it is a low consensus document. The criteria we applied in selecting high and low consensus documents is as follows:

- For a relevant document at the high consensus level, its probability of relevance has to be \( \geq 0.8 \), while it has to be \( \leq 0.2 \) for a non-relevant document at the same level of consensus.
- The probability of relevance of a low consensus level document has to be \( \geq 0.4 \) and \( \leq 0.6 \).

The probability of relevance is equal to the fraction of participants that judged a document to be relevant. To make this point clearer, let us take the following scenario. Suppose document C was judged by 10 assessors. Nine of them judged it as not relevant and just one of them said the document is relevant. Then, the probability of relevance for document C is not exactly 0 but it is 0.1. However, if all the 10 assessors judged it as not relevant then its probability of relevance would be 0.

When we chose documents, we considered documents which were judged by at least six participants or more and divided the documents into two levels of consensus: high and low. We selected 9 documents for each search topic in the study. 6 documents at the high consensus level and 3 documents at the low level of consensus. Of the 6 documents at the high consensus level, 3 are relevant and 3 are not relevant per NIST. Table 1 shows the number of documents at each consensus level and their corresponding NIST qrel scores.

The high consensus documents are those that we are quite certain that secondary assessors will agree on their judgments. The low consensus documents are those that we expect secondary assessors to differ on their judgments.

We are interested in user behavior with top ranked results. To find top ranked documents, we used the parameters given by Cormack et al. [6] to calculate the reciprocal rank fusion (RRF) for each document that we chose in the study. To compute the RRF score, we used all the runs submitted to the TREC 2005 Robust track. When a document is retrieved near the top of most ranked lists, this means its RRF is high, while it will be low if it is not among the top retrieved documents. When choosing between similar documents, we chose documents that had higher RRF scores.

We used the Latin Square method to balance the assignment of a topic to a task across different participants [2]. Our main aim was to reduce any effect of the order of topics in which the participants completed the study. We also randomized the order of documents in each task for each participant.

<table>
<thead>
<tr>
<th>NIST qrel</th>
<th>NR=0</th>
<th>R=1</th>
<th>HR=2</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Consensus</td>
<td>Non-Relevant</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>High Consensus</td>
<td>Relevant</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>Low Consensus</td>
<td>Non-Rel/Rel</td>
<td>3</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 1: Distribution of Documents in Low and High Consensus Levels and their NIST qrel Scores. NIST qrel Scores 0, 1, and 2 stand for “Not-Relevant”, “Relevant” and “Highly Relevant” respectively.

### 3.3 Study Protocol

The study was divided into two parts. The first part of the study was devoted to training the assessors about the required tasks, and the second one was devoted to performing the assigned search tasks. Our study was held in a private room at the university. The room provided a good level of privacy and quiet for conducting the study.

We recorded not only what was shown to participants on the screen and their interaction with it but also their voices. All the reported results in this paper are based on the transcribed data we collected in the study and the video recordings. The following is detailed information about each part of the study.

#### 3.3.1 Tutorial Part

As it should happen in any study and to produce more reliable data for a think-aloud study, participants must have enough training [3,17]. Therefore, our main goal in the training part was to give them the required training so that they felt confident performing the task part of the study. Moreover, the training part will help them to practice how to verbalize their thoughts. Four documents were given to each participant in this part, and participants were required to judge the document’s relevancy to a given search topic. Two of the documents were relevant to the given search topic, and the other two were not relevant. The task of judging documents for a search topic is referred to as a search task in this paper, and that is different from a task for judging just a single document. This distinction should be clear.

During the tutorial part, participants were told if their answers were correct or not and, in few cases, the researcher made some explanations or gave more details if there was a need. The participants were informed that if any elements
were unclear during the tutorial, they should seek clarification, as during the task part, they would work solo, with no intervention from the researcher.

### 3.3.2 Task Part

In the main part of the study, the participants were left to work on their own. At this part, the only role of the researcher was to encourage the participant to continue speaking when there was a lengthy period of silence. The researcher did so since this is the primary characteristic of the think-aloud protocol, where participants speak while performing the required tasks, thereby allowing for data collection. Each participant in this part performed 4 search tasks. In each search task, he judged 9 documents as either relevant or not relevant. As a result, 36 relevance judgments were made by each participant.

### 3.3.3 User Interface

We designed our user interface to accept only binary judgments, where assessors must judge a document as either relevant or not relevant. We did so since the collection of binary relevance has been the most standard form of relevance judging.

The designed interface showed participants one full document at a time. To ensure that participants remembered the search topic and its description during the task, the title of the search topic and its description remained available on the interface. The participant cannot rejudge the document once he goes to the next one. The judgment that he made is final.

### 3.4 Participants

We started the recruitment process immediately after getting the ethics approval from our university. We used email to recruit participants. We sent an email to the Grad-Mailing List to which all graduate students in the university are subscribed. In total, 14 graduate students participated in the study. We used the first two participants to pilot the study. Therefore, the actual number of participants is 12. However, from the remaining 12 participants, we excluded the data of 3 participants since there was unintended intervention from the researcher with these 3 participants. We provide more detail about this exclusion in subsection 3.5. All reported results are from the remaining 9 participants. The participants were from different academic backgrounds. Six of them were science, technology, engineering, or mathematics students; two of them were arts students; three were from environmental studies; and one was from health studies. The average age was 25 years, minimum was 21, and maximum was 35. Rather than confound the results with English comprehension issues, we required all participants to be native speakers of English.

The majority of the participants stated that they use search engines several times a day and they believe themselves to be experts in finding information on the Internet. None of the them had received training in information retrieval. However, two of them indicated that they had received introductory lessons in accessing library resources and article databases.

### 3.5 Cleaning of Data

During the engagement of the participants in the relevance judgment process, there were three times where they were confused about one of the search topics and asked the researcher directly about it. The researcher answered them by accident since he was not expecting questions from participants during the relevance judgment process. However, in order to obtain clean results that are not biased or impacted by the research facilitator, we identified all the places where the interventions have occurred. Then, we excluded all the data that belong to the participants whom the intervention have occurred with.

It is worth mentioning that all the interventions have occurred with search topic 310 which is about Radio Waves and Brain Cancer. In these interventions, the participants were asking about “car phones”. It appeared that they knew nothing about them and never heard of them. We notice that these participants are under the age of 25 and to them this term is not known. After cleaning all data in the study, we were left with 324 relevance judgments. All our analysis and results are based on them.

### 4. RESULTS

This study is an effort to get a better understanding of secondary assessors’ relevance judgment behavior, to explore the real causes of making different relevance judgments, and to determine if assessors make guesses while judging documents. The think-aloud method is the tool that we used to address our concerns in this paper. We were able to listen directly to what assessors said and thought while they were engaged in performing search tasks.

The following subsections discuss the results of our study in more detail.

#### 4.1 Certainty in Relevance Judgments

When analyzing the transcribed data, which is produced by participants in our study, participants expressed different levels of certainty in their relevance judgments. This is shown by the types of expressions and phrases that we noted in the transcripts. Figure 1 illustrates the expressions of certainty that we found. At the two ends of the spectrum in this figure, there are two entirely different types of certainty. The most uncertain relevance judgments start from the far left of Figure 1, and while we go toward the right on the same figure, the level of certainty increases dramatically. Based on these expressions, we believe that relevance judging is not just a matter of two distinct choices: relevant or non-relevant. Users sometimes make highly certain judgments while, in other cases, the certainty in their relevance judgments is completely absent. Between these two extremes, there are other levels of certainty that we are not able to identify clearly by using common quantitative research methods; however, the think-aloud method helped us to clearly achieve that.

The forthcoming sub-subsections discuss the level of certainty that we have found in the collected data.

##### 4.1.1 Low Certainty Relevance Judgments

Low certainty relevance judgments represent the fraction of judgments, which are produced when participants are confused and do not know what to do. Judgments here are not more than guesses. Participants at this level of certainty believe that the binary relevance grading scale, which the interface offers, does not represent their judgments correctly. However, they have to make a decision; therefore, they just make a random choice.
In the following example, Participant 2 did not know what Attention Deficit Disorder (ADD) is. This lack of knowledge became clear as the participant worked on topic 383 (Mental Illness Drugs). Topic 383 requires the relevance assessor to find the names of drugs used to treat mental illness. From the transcribed data, which we obtained via the think-aloud method, we can note that the lack of knowledge about this term (ADD), led the participant to be confused and as a result, produce a low level certain judgment by saying “I’m not sure.” Here is the complete think-aloud transcript for this example with our comments in parentheses:

Are we over-medicating our kids? (Participant reads the title of the document.) How is this one related to mental illness? Was impulsive, wouldn’t sit down. (Part of document read out loud.) Attention deficit, ADD, Ritalin. (Has found the name of drug.) Is ADD a mental illness? (Asks own question.) Hmm, I don’t think so. (Answers own question.) I’m not sure. (Said as participant marks document non-relevant.) Participant 2

4.1.2 Medium Certainty Relevance Judgments

There is more certainty in participants’ relevance judgments at the medium level than those at low level. However, the main characteristic of this level is the lack of total certainty. In the following example, participant 11 is trying to decide if a document is relevant to topic 436 (Railway Accidents). Topic 436 requires the relevance assessor to find documents that provide data on railway accidents of any sort. The description of topic 436 also provides more detail about what should consider or not when judging documents.

Germany to replace rail chief. It doesn’t sound relevant. So far. But it is talking about the person. More talking about the person. No mention of an accident. Oh, what, they do talk about one crash in 1998. Meant that killed 101 people. And that, talking about how he handled it. It provide some information about the accident but not that much. Ah, it doesn’t really focus on the, the accident. And it doesn’t provide any really damage data to see there is an accident. That killed 101 people. But I don’t know if that makes it relevant enough. Uhm, I think says, it’s a, it says how many people were killed that counts as data and I think that makes it relevant. Participant 11

It is obvious in the above example how it was difficult to the participant to fit what he was reading in the document into the criteria mentioned in the description of the search topic. We can note that the participant was going back and forth between the two mandatory options, relevant or non-relevant. This is caused by the lack of ability to find a good criteria fit. Therefore, producing what we call a medium level relevance judgment was the choice of the participant; however, like the judgments at the low level certainty, judgments at this level were transformed by force into a binary relevance judgments due to the two options that the interface provides.

4.1.3 Certain Relevance Judgments

At the other end of the spectrum in Figure 1, participants had many ways to express that they were completely, or near completely, certain in their judgments.

For example, participant 8 is definite about his judgment. In this example, mentioning the names of the drugs in the document’s title made the participant sure about the decision that he made.

Okay, So the White House seeks to curb. Alright, so this mentions drugs that are used for mental illness right in the title. Definitely relevant. Participant 8

Also, in the following example, Participant 11 is certain about the relevance judgment that he made. It appears from the expressions that he was using it does not sound that it could be relevant that he is not definitely certain (highly certain) about his decision but at least certain.

It doesn’t sound that it could be relevant because it doesn’t talk about an attack in the title. It just talks about bigfoot and no attacks. It is not relevant. Participant 11
4.1.4 Frequency of Certainty Levels

Table 2 shows certainty levels that we have found in the participants’ answers and their frequency in our data. One of the authors assigned a level of certainty for each judgment in the study. As noted in the table, most of the relevance judgments are made with certain to medium certainty. While the low certainty judgments represent a small fraction — 25 of the 324 total relevance judgments (7.71%) — the important observation here is that even secondary assessors attempting to make good judgments, will be forced to guess the relevance of some documents if the only interface action allowed is a relevance judgment.

<table>
<thead>
<tr>
<th>Certainty Level</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Certain</td>
<td>196</td>
</tr>
<tr>
<td>Medium Certain</td>
<td>100</td>
</tr>
<tr>
<td>Low Certain</td>
<td>25</td>
</tr>
<tr>
<td>Unknown</td>
<td>3</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>324</td>
</tr>
</tbody>
</table>

Table 2: Certainty levels and their frequency. Unknown in this table refers to assessors’ relevance judgments that we were not able to categorize under one of our three main certainty levels.

4.2 Making Incorrect Relevance Judgments

We analyzed participants’ incorrect relevance judgments based on high and low consensus levels as well as NIST qrel scores. There was a feeling before analyzing the data that the reasons for making incorrect relevance judgments might differ based on the consensus of the documents (high or low). However, based on our analysis of the transcripts and the recorded videos for the incorrect relevance judgments, we found the reasons that lead participants to make incorrect relevance judgments are the same despite the consensus levels (high or low).

Also, when analyzing the data, we considered not only the consensus levels but also the NIST qrel scores. For example, when a document is considered a high consensus one, we checked also its NIST qrel score. NIST used a graded relevance scale: 0, 1, or 2 that represents non-relevant, relevant, and highly relevant respectively. Table 1 shows the number of documents at each level of consensus and their corresponding NIST qrel scores.

4.2.1 Categories of Incorrect Relevance Judgments

Even though quantitative research methods are able to provide us with a good level of understanding of assessors’ (users’) behavior, qualitative research methods (in our study represented by the think-aloud protocol) are an excellent way to understand why secondary assessors make incorrect relevance judgments. One of the authors analyzed the think-aloud transcripts and the video recordings and found that incorrect relevance judgments can be divided into four categories:

**Search Topic:** Under this category, the participant is not able to apply the given search topic to the document. Sometimes the participants found it hard or unclear as to how to apply the search topic. We call this *Difficulty in Applying Search Topic.* Difficulty in applying search topic occurs when the assessor finds what he thinks relevant information. He rechecks the search topic to make sure that this information fits into the criteria mentioned in the description of the search topic.

China sets railway safety. OK, so does it have information on accidents? Yes, Human deaths. OK, And then, does it talk about prevention or safety? So documents that discuss railroading in general, new rail lines, new technology for safety, and safety and accident prevention are not relevant, unless an actual accident is described. In the example below, the assessor found some data about human deaths. He continued reading the document and found some data about safety. Therefore, he believes that since the document talks about safety and there is also data about human deaths then it is relevant. However, this document is judged by NIST assessor as not relevant and is among the high consensus documents (Not Relevant group) in our experiment.

**Document:** Even though the participant understands the search topic and knows what he is looking for, it is hard
to process the document and find the relevant content. We call this *Difficulty in Processing Document*. The following example shows how the assessor’s inability to process the document caused him to judge it incorrectly. From the collected data, we note that the assessor in this example knows what he should look for in the document. However, he failed to find that the documents says “Candy wrappers were found in the tent . . .”, which is relevant. Table 5 shows the topic description that notes that mention of such causes will make a document relevant. This document was judged as a highly relevant document by the NIST assessor. However, the participant missed the relevant content in the document and judged it as non-relevant.

Well, this article actually does not seem to be relevant because it does not discuss the frequency. It is just talking about one isolated incident and does not really talk about any causes of the attacks. And there is no scientific, speculative or anything like that (Said as the participant marks the document as non-relevant.) **Participant 12**

**Assessor:** Under this category, the participant lacks the required knowledge, which prevents him from judging the documents correctly, or he lacks concentration. The following example illustrates how the assessor’s lack of knowledge caused the incorrect judgment. The assessor in this example was working on topic 310 (Radio Waves and Brain Cancer) which requires that relevance assessor to find documents that provide evidence that radio waves from radio towers or cell phones affect brain cancer occurrences. As it is clear in the example, the assessor does not understand that microwave radiation is a form of electromagnetic radiation and that caused him to stop reading and judged the document as not relevant.

So, I am scanning the article for cancer. Uhm, that says microwave. Radiation not radio towers. So, I would say this one is not relevant. **Participant 10**

**NIST Assessors’ Mistakes:** This category is different from the above three. It occurred just with only one document and for a specific search topic. As we mentioned in this subsection, we compared our participants’ relevance judgments against the judgments made by NIST assessors. Even though NIST assessors are well-trained, primary assessors, they still might judge documents incorrectly. We found that a non-relevant document to NIST assessor was judged several times as relevant document by our participants. The types of answers and proofs that participants provided us with led us to recheck that document carefully. After analyzing the whole document carefully, we believe that the NIST assessor was wrong in judging this document as a non-relevant one for the Mental Illness Drugs topic (Topic 383). The document in question has a docno of NYT19991214.0159.

Topic 383 requires the relevance assessor to find the names of drugs used to treat mental illness. The following example illustrates the judgment that was made by one of the participants in the study in regard to the above mentioned document:

But then it mention specific, Prozac and Zoloft and depression (Participant has found a specific drug name). So it’s relevant. **Participant 10**
negatives, we need to build relevance assessing systems that assist assessors in finding relevant material in documents.

Difficulty in applying the search topic can result in judging issues during the search of a document or during the decision concerning topic fit. For example, if an assessor does not understand a topic, this can result in either missing relevant material or in the misidentification of material. Once the assessor finds potentially relevant material, difficulty with the search topic can again produce differences as the assessor makes their final relevance decision. In the case of documents that had high consensus, the few errors caused by the search topic seem divided between false positives and negatives. In the case of low consensus documents, search topic errors were much more frequent and tended to result in false positives. In general, secondary assessors seem to lack understanding of detailed topic criteria that leads to false positive errors.

4.2.2 Assessor Causes of Differences

We found the following specific assessor causes of judgment differences:

- **Trouble Understanding the Search Topic**: This is caused when the participant does not fully understand the search topic or part of it. Here are examples from the think-aloud transcripts where the participants misunderstand what a generically named drug is:

  Specific or generic type of drug. Uhm. So, I guess that’s a generic type of drug. Antipsychotic. Uhm. So I guess I would say it’s relevant. Participant 10

  It talks about antipsychotic drugs. I guess that’s a generic type of drug. I think this is relevant. Participant 8

- **Lack of Familiarity/Lack of Knowledge**: In this reason, the participant is not able to figure out if the document is relevant or not due to a lack of familiarity with the given search topic or the lack of familiarity with some words or phrases in the document.

  Not sure of that depression is a mental illness. So, I am going to think about that. Okay, so Ritalin is called stimulant because it belongs to a family of drugs that stimulate the central nervous system. Okay, ahh, parent’s alertness and ability to pay attention. Oh, increases a person’s alertness. So, I don’t think that is about the mental illness. I am not sure what qualifies a mental illness. I am thinking. Participant 5

- **Lack of concentration**: In very rare cases in our collected data, lack of concentration, which we describes as a kind of user behavior, was noticed. Lack of concentration was the cause of simple mistakes. Participants judged documents as relevant while those documents were entirely unrelated to the search topic.

  In the below example, Participant 2 judged the document as relevant even though it discussed grizzly bears and not black bears. He did not notice this, and attempted to fit the document into the criteria mentioned in the description of the search topic. This is clear when the participant said *See if there is any recommendations and “See if there is any ways to control it”*. The participant said that he knows generally about the search topic Black Bears Attacks in the pre-task questionnaire; and grizzly bears are different from black bears.

  Behind conflict over new grizzly program, an endangered species war. Four weathered men stand around, dropped tailgate. Fruits leading cave. This is the kitchen that where the food is. Okay. Okay it was talking about food that’s a cause. Threatened animals. See if there is any recommendations. See if there’s any ways to control it. Oh. Okay. That’s good. Participant 2

4.2.3 Assessor Decision Making

In examining the collected data for the cases where assessors made judgments that differed from the primary assessor, we found the assessors giving the following reasons for their decisions:

- **Insufficient Information**: Sometimes, participants might think that the document does not have enough information and this will cause incorrect judgments. The following is a think-aloud data example from the transcripts:

  So this is not relevant at all, we’re not, it’s just too short, we’re not getting enough information about anything. Participant 4

- **Presence or Absence of Specific Evidence**: Participants affected by this factor cannot find evidence (specific information) or they think they find the evidence. Here is an example:

  So it’s not relevant because it didn’t provide any evidence of that connection yet. Just about certain studies. Participant 5

- **Lack of Topicality**: Participants incorrectly think that the document is off-topic. What follows is an example of this:

  It’s not talking about an incidents or anything. It’s just talking about some crazy lady and why you should keep bears. Participant 4

- **Absence of Keywords**: Some participants were looking mainly for keywords in the documents. If key words were not there, that meant the document was not relevant to them. Here is an example:

4.3 Short vs. Long Search Topics

In examining the differences in relevance judgments at a per-topic level, we noticed a difference in the rates at which participants misidentified relevant documents as non-relevant and vice versa. As shown in Table 6, the false negative rate (FNR), which is defined as the fraction of relevant documents incorrectly judged as non-relevant, is higher when the description of the search topic is long. Our interpretation for this kind of behavior is that the description provides more details and criteria for relevance, which sometimes hard to remember by assessors or might mislead them. Moreover, participants when they encounter this type of description seem to focus on some parts of the description during the relevance judgment process. For example, the description of topic 336, which is about Black Bear Attacks, is shown in Table 5.

If we look at the description of topic 336 in Table 5, we note that the first, the third and the last sentences describe what should be considered when an assessor looks for relevant information. These three sentences are as shown below and we start with the first sentence:

A relevant document would discuss the frequency of vicious black bear attacks worldwide and the possible causes for this savage behavior.

the third sentence is as the following:

Relevant documents would include the aforementioned causes as well as speculation preferably from the scientific community as to other possible causes of vicious attacks by black bears.

finally, the last sentence says the following:

A relevant document would also detail steps taken or new methods devised by wildlife officials to control and/or modify the savageness of the black bear.

We can see clearly that every sentence in the above three sentences is adding more data to consider in the relevance judgment process. We found from the transcribed data that participants are not able to accumulate all of these points in the description. The following examples are taken from the transcribed data:

So, boy is attacked by bear at scout camp. So, let’s look to see if there is anything about frequency of attacks or causes. Talks about the savageness of the black bear.

In the above example, participant 5 was also looking just about scaring people. They are not actual attacks. Participant 5

So this just seems like a bear attacking a person but a bear being in a property. So, uh, and attacking the goats. So, I don’t know. It is relevant yet. Ahh, maybe an about [...] The bear showing at a barbeque. So, I don’t know if that counts as an attack. Just scare some people. Yeah, this is more about [...] Ahh, oops! So, I am going to say not relevant because it is just about scaring people. They are not actual attacks. Participant 5

In the above example, participant 5 was also looking just for the causes of the attacks and not considering the other parts of the description which might lead him/her to find the relevant information.

On the other hand, we found that the false positive rate (FPR), which defined as the fraction of non-relevant documents incorrectly judged as relevant, for search topics that have short descriptions is higher than those which have longer descriptions as illustrated in Table 6. Since there is just little information in the description about what to consider when looking for relevant documents, this may lead participants to be more liberal in this regard and to consider the non-relevant documents as relevant. Table 7 represents an example of a short topic’s description.

5. CONCLUSION

Based on the results presented in this paper, assessors’ relevance judgments are associated with different levels of certainty ranging from low certainty to high certainty. Assessors that lack certainty in their judgments are bound to produce judgments that differ. We also found that the search...
topic, the document, and the assessor can all interact and be the cause of differing judgments. Assessors may have trouble applying a search topic’s criteria to a given document, or assessors themselves may lack required knowledge, and sometimes certain documents make finding relevant material difficult. Finally, we found preliminary evidence that the number of details that is given in the description of the search topic appears to affect assessors’ judging behavior. These findings point to several user interface changes that we intend to investigate in future work. We believe that relevance assessing systems should be designed to not only collect relevance judgments, but that these systems should also provide means for the efficient collection of judgment certainty and judgment rationale. While the collection of this information may not immediately produce better judgments, the collection of this information should help identify issues with secondary assessors’ interaction with search topics and documents.

6. ACKNOWLEDGMENTS

This work was supported in part by King Saud bin Abdulaziz University for Health Sciences (KSAU-HS), in part by Natural Sciences and Engineering Research Council of Canada (NSERC), in part by a IIiX’14 Student Travel Grant, and in part by the University of Waterloo.

7. REFERENCES