Presentation Ordering Effects On Assessor Agreement

Tadele T. Damessie
RMIT University
tadeletedla.damessie@rmit.edu.au

J. Shane Culpepper
RMIT University
shane.culpepper@rmit.edu.au

Jaewon Kim
RMIT University
jaewon.kim@rmit.edu.au

Falk Scholer
RMIT University
falk.scholer@rmit.edu.au

ABSTRACT

Consistency of relevance judgments is a vital issue for the construction of test collections in information retrieval. As human relevance assessments are costly, and large collections can contain many documents of varying relevance, collecting reliable judgments is a critical component to building reusable test collections. We explore the impact of document presentation order on human relevance assessments. Our primary goal is to determine if assessor disagreement can be minimized through the order in which documents are presented to assessors. To achieve this goal, we compare two commonly used presentation orderings with a new ordering designed to aid assessors to more easily discriminate between relevant and non-relevant documents. By carefully controlling the presentation ordering, assessors can more quickly converge on a consistent notion of relevance during the assessment exercise, leading to higher overall judging agreement. In addition, important interactions between presentation ordering and topic difficulty on assessor agreement are highlighted. Our findings suggest that document presentation order does indeed have a substantial impact on assessor agreement , and that our new ordering is more robust than previous approaches across a variety of different topic types.

ACM Reference format:
Tadele T. Damessie, J. Shane Culpepper, Jaewon Kim, and Falk Scholer. 2018. Presentation Ordering Effects On Assessor Agreement. In *Proceedings of The 27th ACM International Conference on Information and Knowledge Management, Torino, Italy, October 22–26, 2018 (CIKM '18),* 10 pages.
DOI: 10.1145/3269206.3271750

1 INTRODUCTION

Test collections are at the heart of information retrieval (IR) evaluation, and the Cranfield projects which started in early 1960*s* provided a foundation for the methodologies used today to evaluate IR systems [9, 18, 32]. Many highly influential IR evaluation campaigns have built on these foundations, including TREC [48], NTCIR [23], CLEF [5] and FIRE [33]. Test collections are the most widely-used approach for evaluating the effectiveness of information retrieval systems [36], and human relevance assessments – indicating the responsiveness of answer items to a search topic – are the most resource-intensive component when creating such evaluation testbeds, requiring time, cognitive effort, and often money. The introduction of graded relevance levels in most recent test collections [19] has further increased the costs, as collecting credible and consistent relevance judgments using human assessors is impacted by the greater range of possible subjective interpretations of relevance introduced by graded relevance levels [46].

The introduction of crowdsourced relevance judgments has provided a relatively low-cost and fast method of creating new relevance assessments, but with an associated risk that such assessments may be of lower quality, and with lower agreement between assessors, if the judgments are not collected carefully [17, 26, 40]. Various factors have been shown to affect how assessors judge documents, including topic familiarity [4], topic knowledge [31] and the degree of document relevance that is encountered early in the judging process [38].

Recent work has demonstrated that the level of agreement between assessors can be used to gauge the quality of the relevance judgments [12]. In this work, we investigate two key variables – presentation ordering and topic difficulty – and study their impact on assessor agreement. A user's understanding of a pre-defined topic can have a strong effect on their ability to distinguish between relevant and non-relevant documents, and *priming* is a well-known technique which can reduce *relevance drift* as assessors judge documents [38, 45]. Leveraging these concepts, we propose a new presentation ordering technique that interleaves the most and least likely relevant documents; we call this approach *Interleaved Likelihood of Relevance* (ILR). Our study aims to determine if assessor disagreement can be directly minimized through the order in which documents are presented to assessors. The key idea is to maximize the difference between relevant and non-relevant documents presented to assessors, so they can more quickly converge on a consistent notion of relevance during the assessment exercise. We show that ILR is more resilient to variance in topic difficulty, and leads to higher overall agreement, than two more common presentation orderings which are widely used to gather relevance judgments in building test collections – *Decreasing Likelihood of Relevance* (DLR) and *Random Likelihood of Relevance* (RLR).

The following research questions are investigated:

RQ1: *What is the relationship between user-based and system-based notions of topic difficulty?*

RQ2: *How does presentation ordering affect inter-rater agreement when judging the relevance of documents?*

CIKM '18, Torino, Italy

978-1-4503-6014-2/18/10…$15.00
DOI: 10.1145/3269206.3271750

RQ3: *How do system- and user-based notions of difficulty for a topic interact with presentation order when judging the relevance of documents?*

Contributions. In order to answer these questions, we conducted a large user study of 96 users across 8 topics of varying difficulty. We adopt a *Gold Benchmarking* experimental setup, where the relevance assessment task is set up independently of any existing relevance information (but existing relevance judgments can be used to validate the quality of the setup, if available). We analyze inter-rater agreement and interactions between topic difficulty and order of document presentation to address our research questions. Finally, we introduce a new method of document presentation ordering (ILR) which maximizes agreement, consistency and the quality of relevance judgments in building reusable test collections.

2 RELATED WORK

Human relevance judgments are crucial in the formation of test collections for IR system evaluation. The judgments gathered from human assessors are subjective and likely to result in some level of disagreement [11, 37, 42]. The implications of this are two-fold. First, inter-rater agreement may not be high [25], and second, effectiveness scores from test collections using judgments from a single assessor may be inconsistent [36]. As a result, the magnitude of system effectiveness scores might change if results are computed using judgments from different human assessors; which might impact relative system orderings during benchmark comparisons [47].

According to Bailey et al. [4], human assessors creating relevance judgments fall into three classes: gold, silver or bronze. Gold standard assessors are topic authors and subject experts at the same time, and are sources of the highest quality judgments. This class of assessment has been common in TREC ad hoc tasks, for example, but is typically difficult or costly to obtain when creating new test collections. Silver assessors are domain experts who are not topic originators. Bronze assessors are those who are neither topic originators nor subject experts, as might typically be the case when employing crowdworkers to conduct relevance assessments.

Identifying and controlling the influencing factors of relevance assessment has an important role in the consistency of judgments collected from two or more assessors; and hence a reusable test collection. Eisenberg and Barry [15], Huang and Wang [22] investigated the relationship between document presentation order and relevance judgment. They concluded that relevance scores are overestimated when using increasing relevance document presentation ordering; and underestimated when documents are presented in a decreasing relevance level order.

Research on relevance assessment exercises has found that agreement can also be influenced by topic difficulty [2] in a crowdsourcing platform. The study asked assessors to rate topics on a scale of 1 (easy) to 5 (very difficulty), and found an inverse correlation between agreement and difficulty. This work differs from ours in several ways. First, our study selects topics of various system and user difficulty levels. Second our main focus is the interplay between topic difficulty, presentation order and agreement, and not about the minimizing the cost of gathering relevance assessments with crowdsourcing platforms. Third, our experiment is a lab-based user study, together with a new crowd-sourced study to investigate user-based notions of topic difficulty. Finally, in our study, the number of relevant documents presented to assessors is estimated using the distribution of relevance in the collection.

Inter-rater agreement and inter-rater reliability are related concepts with a few technical differences [16, 29]. Agreement refers the degree which ratings of two or more judges are identical, whereas reliability relates to rater concordance in the relative ordering of subjects under investigation. Higher agreement means different judges assign exactly the same value for each judgment; higher reliability on the other hand indicates the relative "ordering" of all judgments between all of the assessors are consistent. These two concepts are commonly used to quantify the quality of judgments in crowdsourcing experiments [17, 26, 40]. Kazai [26] used agreement to clean assessments using gold standard judgments, and found that the quality of relevance judgments was improved. Grady and Lease [17] suggested that it is more efficient and effective to collect fewer assessments when agreement is higher.

Scales of ratings used can impact both agreement and reliability. For example, a binary scale is more likely to produce higher agreement due to the probability of ratings being similar; but the same scale might have lower reliability because a mismatch yields an inverted relative ordering unlike ratings on a scale of more than two choices [27]. Using too many levels in a rating scale could overwhelm assessors and lead to lower inter-rater agreement [17]. While the optimal number of levels, and the meaning assigned to them, is an open issue [7], we adopt the widely-used 4-level relevance scale as defined by Sormunen [41] and Kekäläinen [27].

In this research, managing the order in which documents are presented to assessors and the interaction of document presentation ordering and topic difficulty will be examined as a proposal to maximize and improve consistency of relevance assessment.

3 METHODOLOGY

Evaluating the effectiveness of an IR system often relies on the construction of a test collection. A test collection has queries, documents and a set of relevance judgments indicating which documents are relevant to which query. The relevance judgments are typically created manually by human assessors.

In a relevance assessment task, bronze class assessors (i.e. those who are not the creators of the test topic, nor subject experts in the topic's domain, such as would for example be the typical case where relevance judgments are crowdsourced), together with some existing gold standard judgments can be used in one of two ways. The first alternative is to use gold judgments to inform the design of a relevance assessment task. That is, the selection of documents and topics for an experiment are based on the information available from existing gold assessments [2, 13, 14]. We refer to this as *Gold Guided Design*.

A second alternative is to set up a relevance assessment task independent of the information available in the gold assessment; and to potentially use gold assessments (if available) to benchmark various aspects of the assessment results [35, 47]. We refer to this as *Gold Benchmarking*. Our study is based on the later alternative.

Damessie et al. [14] investigated the effect of document ordering and topic difficulty on assessor agreement using a Gold Guided Design. The study experimented with two commonly used document ordering techniques in IR evaluation campaigns: *relevance*

order [34] and *document identifier order* [48]. With relevance ordering, documents are presented to assessors from highest to least relevant for a topic; and in document identifier order, documents are ordered using the TREC assigned document identifier, from here on referred to as docID. Their study found higher inter-rater agreement for docID order than relevance order. The higher agreement for docID order was explained with what they called the "surprise effect". That is, documents presented in decreasing relevance order are subject to relevance underestimation due to highly relevant documents being seen early in the assessment list [15, 22]. For docID order, relevant documents are spread across the assessment list and interspersed with non-relevant documents, creating the "surprise effect" and allowing assessors to more easily distinguish relevant documents from non-relevant ones.

There were two major caveats of the Damessie et al. [14] study. First, when creating a new test collection, relevance judgments are rarely available for pre-sorting documents prior to presenting them to assessment. Second, while docID order may create a "surprise effect" as reported in the study, this is not a given. Figure 1 shows the results of a simulation carried out using the TREC 7 and 8 collections, where thirty documents were sampled randomly from 5 topics, and then presented in either docID (D) or random (R) order, shown as pairs of rows. As can be seen, relevant documents often cluster together in docID order. This is due to the document collections from TREC 7 and 8 being composed of several different sub-collections from different NewsWire sources, with each document being named with a prefix string that identifies its sub-collection of origin. Relevant documents can often occur predominantly in a single sub-collection, and after down-sampling to a smaller number of documents, and then sorting by docID, relevant documents from the same sub-collection are more likely to co-occur. Damessie et al. [14] also found that topic difficulty can impact inter-rater agreement in addition to document ordering. System easy topics (topics that have a high AP score for many different systems) tend to have a higher overall inter-rater agreement than system difficult topics. However, their study was small, did not compare interaction effects between topic difficulty and presentation ordering, used a Gold Guided Design, and did not provide any insights into how to operationalize the "surprise effect".

As this work is closest to our own, we leverage their framework and lessons learned to further investigate how document presentation order influences agreement. To this end, we set up an experiment with three different document presentation orderings. Two common orderings are used as a baseline – *Decreasing Likelihood of Relevance* (DLR) and *Random Likelihood of Relevance* (RLR), which are both adaptations of commonly used during test collection construction exercises at international IR evaluation campaigns [34, 48], but without unexpected consequences of document clustering observed in docID ordering as we are using the same test collection. In addition, we propose a new presentation ordering called *Interleaved Likelihood of Relevance* (ILR) where a careful combination of those documents that are most and least likely to be relevant are interleaved and presented to the user.

Note that our investigation is not based on Gold Guided Design principles. Rather, we explore the notion of likelihood of relevance. The likelihood ordering which approximates a decreasing relevance

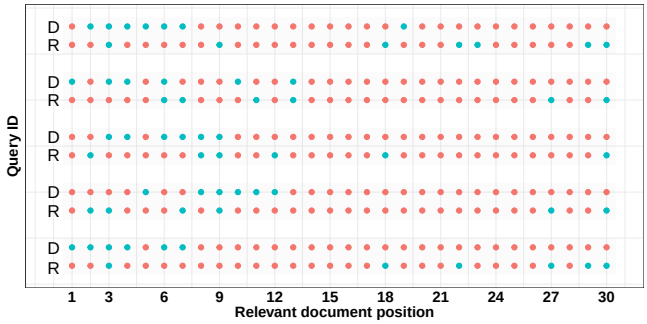

Figure 1: Position of relevant documents simulated for randomly selected 5 topics for a sample of 30 documents ordered using TREC identifer (D) and random method (R). Red shows the position of non-relevant documents and green shows the position of relevant documents. The gray shaded region highlights documents in a topic ordering using both methods.

order (DLR) is derived directly from NTCIRPOOL [35], a pseudo-relevance approach based on the number or runs that return a document (the higher the better), and the sum of document rank positions in ranked lists (the lower the rank, the better). RLR is produced by randomly shuffling the documents from the DLR list. Our new approach, ILR, is motivated directly by the surprise effect reported by Damessie et al. [14].

As previously explained, we also consider the influence of topic difficulty, from both a system and a user perspective. Since these factors will influence the choices of documents for our main study, we next describe a crowdsourcing study on the issue of measuring user-perceived topic difficulty.

3.1 Crowdsourcing Study: User Difficulty

Our first user study was conducted via crowdsourcing, using the CrowdFlower platform.[1] The goal of the used study was two-fold. First, we wanted to better understand the relationship between *user difficulty* (a human user's perception of how difficult a topic is) and *system difficulty* (a measure of how difficult it is for the system to supply a good ranked results list). We also wanted to select appropriate topics for a controlled user study (presented in the next section) and therefore we needed to identify topics of varying difficulty for the experiment. The CrowdFlower study made use of a subset of topics and documents from a collection created by Sormunen [41], where a set of 41 topics from the TREC-7 and TREC-8 test collections were re-judged on a 4-level relevance scale. We set out to select a subset of 8 topics from this set, as this corresponded to our available resources for the subsequent lab-based user study. Selection was carried out on the basis of both system and user difficulty.

System difficulty. System difficulty aims to capture the notion of how hard it is for a retrieval system to supply a good ranked results list in response to a topic. We make use of Average Average Precision (AAP), calculated using the average of Average Precision (AP) values for a topic across all systems [24, 30]:

[1] www.crowdflower.com

Table 1: Questions used for estimating topic difficulty from a user perspective.

No.	Question
1.	How interested are you to learn more about the topic?
2.	How much do you already know about the topic?
3.	How clear is the information need for the topic?
4.	How difficult do you think it will be to search for information for the topic?
5.	How difficult do you think it will be to determine the relevance of documents for the topic?
6.	Overall how difficult do you think the topic is?

$$AAP(t_j) = \frac{1}{m}\sum_{i=1}^{m} AP(s_i, t_j).$$

Note that AAP is the mean AP for a single topic for many different systems, while MAP is the mean AP across multiple topics for a single system. In our experiments, AAP was calculated for the topics of the 2004 Robust track (which included topics from the TREC-7 and TREC-8 test collections with dual binary and 3-level ordinal relevance judgments) for the 110 runs that participated in the track. Following the approach by Carterette et al. [8] to classify topics as system easy (*SE*), system medium (*SM*) and system hard (*SH*), we split topics into three classes based on AAP scores, leading to a classification of the 41 topics into: 12(29%) *SE* category, with AAP scores in the range (0.3, 1]; 13(32%) *SM* category, with AAP scores in the range [0.3, 0.196); and 16(39%) *SH* category, with AAP scores in the range (0.196, 0]. The difference in the number of topics in the classes is due to some topics having the same AAP score.

User difficulty. We estimated *user topic difficulty* using the 6 questions shown in Table 1, similar to those proposed by Crescenzi et al. [10]. Users were asked to supply answers to these questions on a 5-point Likert scale. Each of the 41 topics was crowdsourced and received 10 assessments, five in each of two scenarios. In the first, workers rated the questions after being shown only the TREC topic statements. In the second case, workers were additionally shown two documents – one highly relevant, and one non-relevant – based on the original Sormunen relevance judgments.

Cronbach's α [44] was used to analyze the internal consistency between question items measuring topic difficulty: $Q4$, $Q5$ and $Q6$. Cronbach's α scores are between 0 and 1, and values above 0.7 are usually considered acceptable to validate consistency between items. Structural equation modeling (*SEM*) is a compatible measure of further analysis when an analysis shifts from individual items to composites [3]. Hence, we used *SEM* with a scaling factor to form a composite of the individual question items measuring the construct factor user difficulty.

The topics were then sorted using the composite of user difficulty – the smaller the value, the easier the users think the topic is. Finally, the topics were split into 3 equal sized classes: 14 as user easy *UE*; 13 as user medium *UM*; and 14 as user hard *UH*. The results and analysis of this modeling are presented in Section 4.1.

Table 2: Topics and the intersection of their user and system difficulty classes; the class label is a combination of U (User), S (System), E (Easy), M (Medium) and H (Hard).

Topic ID	Query	Class
#364	`rabies`	*UE & SE*
#420	`carbon monoxide poisoning`	*UE & SE*
#393	`mercy killing`	*UE & SH*
#442	`heroic acts`	*UE & SH*
#385	`hybrid fuel cars`	*UE & SM*
#400	`amazon rain forest`	*UH & SE*
#416	`three gorges project`	*UH & SM*
#440	`child labor`	*UH & SH*

Final Topic Selection. The final set of 8 topics were selected using the intersection of the system difficulty and user difficulty categories. Table 2 shows topics and their classes with respect to user self rating and AAP score.

Using the 8 topics, the study will analyze if the concept of user topic difficulty is shared between CrowdFlower and lab user study groups; and can be measured uniformly using a set of self rating scales between the two groups in order to resolve RQ1.

3.2 Lab Study: Relevance Judging and Presentation Order

To investigate the influence of presentation order when judging documents, as well as the interaction with system and user topic difficulty (RQ2 and RQ3), we carried out a lab-based user study.

Participants and Study Design. A total of 96 participants took part in the user study, which involved making document relevance assessments. 8 participants abandoned the experiment at different stages of the study, and 4 participants used an invalid age (indicating that they were over 60 when this clearly was not true, casting doubt on the quality of their participation); this data was excluded from the final analysis. The assessments from the remaining 96 assessors (39 female, 56 male and one who preferred not to respond) constituted the analyzed data. Assessors were between the ages of 21 and 39 (mean = 30.01 and SD = 5.02), and 41% of the participants were undergraduate students at RMIT University.

Using a 5-point Likert-scale question "I use a search engine like Google, Bing or Yahoo everyday" (1: strongly disagree to 5: strongly agree), most participants (92) classified themselves as regular users of web search engines (mean = 4.59 and SD = 0.642) and for the question "I am good at finding information using a search engine like Google, Bing or Yahoo" (1 strongly disagree to 5 strongly agree), most participants (87) consider themselves good at finding information using search engines (mean = 4.46 and SD = 0.72). The subjects participated voluntarily in the experiment, and were compensated for attending with an AU$20 gift voucher.

In this experiment, a between-subjects design was adopted due to the expected runtime for each topic. To investigate the effect of the research variable document presentation order (DLR, RLR, and ILR), each participant completed a subset of two out of the set of

eight possible topics (selected as described previously). Each topic requires making 30 document judgments. Since timing and fatigue is known to impact user study results [28], estimated participation time was kept to 1 hour, and each assessor was asked to complete 2 topics, giving an average of one minute per document. The presentation sequence for the topic and document ordering variables were counter-balanced across the participants to minimize potential experimental ordering effects [28]; across all participants, every topic was evenly shown with the three different document orderings.

All participants spent around 60 minutes in a controlled laboratory setting to complete their entire task. A script detailing the aim and process of the relevance assessment task was prepared and read out to each individual participant. After participants agreed and signed the consent form, they proceeded to a training exercise. The training exercise is similar to the actual assessment task where a search topic and two documents are displayed, with one document being shown at a time and taking up a full page. Participants were then presented with a pre-experiment questionnaire gathering basic demographic information and familiarity with web search. Following the pre-experiment questionnaire, participants proceeded to the main task. Here, each page of an assessment task displays a topic, a document, and radio buttons of the relevance grades available for assessors to rate documents. A document is the only variable that changes on a page. Following the main assessment of 30 unique documents for a topic, a post-hoc questionnaire about the topic just completed is presented to ask about their interest, knowledge, information need clarity, the difficulty to search information, difficulty to determine document relevance, and overall topic difficulty, each using a 5-point Likert-scale (1: strongly disagree to 5: strongly agree). Assessors were then afforded a 10 minute break before proceeding to a second topic, with a similar procedure, except that the pre-experiment questionnaire was presented only once per assessor.

Topics and dataset. The dataset used in the lab user study is the same as that of the dataset used in the crowdsourcing experiment, with topics selected as detailed in Section 3.1. Topics and the corresponding system and user difficulty categories that they fall into are shown in Table 3. Eight were ultimately selected from these classes for the experiment, as detailed in Section 4.1.

Document Sampling. Participants were asked to judge the relevance of 30 documents for each topic. Therefore, 30 documents from the TREC 7 and 8 test collections were sampled for each of the 8 selected topics. The documents were sampled from document pools, formed using *contributing runs* from the particular test collection. A run was considered to be a contributing run if all of the the top 100 documents of the run have explicit judgments in the official TREC relevance files. A total of 103 and 129 runs were submitted to TREC 7 and TREC 8 respectively, of which 51 in TREC 7 and 60 in TREC 8 passed our *contributing run* filter.

A pool was formed and sorted using the NTCIRPOOL [35] approach. Here, documents are first sorted based on a count of the number of contributing runs that returned the document, in decreasing order. Ties are resolved based on the sum of the rank positions at which documents were retrieved in the runs, in increasing order. Table 5 shows the total number of documents in the pool and number of unique documents for each of the topics used in our study.

Table 3: Topic pools based on the intersection of users rating and AAP score. * indicates classes from which topics are not drawn for our user study.

Class	Topic ID	Total
$UE \cap SE$	#364, #351, #420, #392	4
$UE \cap SH$	#372, #427, #445, #388, #393, #442	6
$UE \cap SM$	#407, #385, #408, #428	4
$UH \cap SE$	#365, #400, #396	3
$UH \cap SH$	#355, #387, #362, #440, #378, #437	6
$UH \cap SM$	#402, #416, #373, #358, #353	5
$UM \cap SE^*$	#410, #403, #431, #415, #418	5
$UM \cap SH^*$	#405, #399, #421, #448	4
$UM \cap SM^*$	#377, #360, #414, #384	4
	Total	41

To obtain a set of 30 documents for each topic, the top 5 and bottom 5 documents are taken from the fully sorted list for that topic. The remaining 20 documents were selected by sampling at regular intervals from the remaining $N - 10$ documents in the lists. The top and bottom 5 documents were included with the aim of making a number of "best" and "worst" documents available for each topic.

Document ordering. Recall that we aim to explore three possible orderings: expected decreasing likelihood of relevance (DLR) and random likelihood of relevance (RLR) – two commonly used document ordering methods in human relevance assessment exercises – and ILR, our proposed approach that maximizes a surprise factor.

We operationalize surprise by estimating the "true" relevance distribution for a set of documents to be judged. Then the subset of most likely to be relevant is interleaved with the subset of most likely to be non-relevant (from most likely to least likely) in order to maximize the surprise effect. We hypothesize that this ordering should significantly increase agreement across assessors. In detail, the ILR method is a three stage process (see Figure 2) where:

- In stage *I*, the assessment list is ordered using DLR order which is produced by NTCIRPOOL. Note that alternative techniques could be used to derive this ordering. We use NTCIRPOOL as this approach has been shown to be correlated with actual document relevance [35], and for reproducibility purposes.
- In stage *II*, the DLR list is divided into equally sized blocks. One of the blocks contains documents which are expected to be relevant. An estimation of proportions of relevance is obtained by considering data external to the set of 8 topics that were selected for this study. For example, in TREC 7, 6% of the documents were relevant; assuming the same percentage, for a topic drawn from this collection that has 961 unique documents in a pool, around 58 documents would be expected to be relevant. When scaled to 30 documents being judged, we expect ≈ 6 of them to

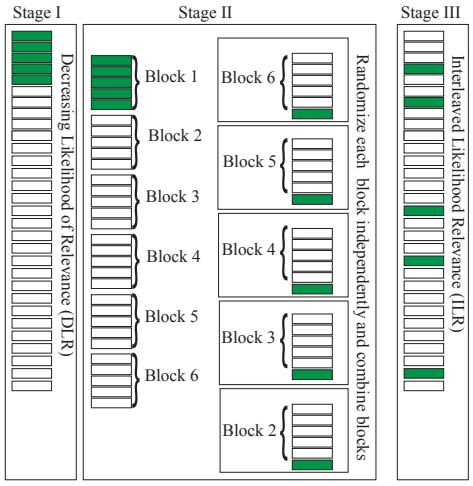

Figure 2: Interleaved ordering as a three stage process. The blocks highlighted in green indicate documents which are most likely to be relevant. The number of documents selected is based on an estimate of how many documents are likely to be relevant on average per topic in the collection.

be relevant. The top block which contains the "most likely to be relevant" documents is used to insert a relevant document into each of the remaining blocks, that are then ordered bottom up. For example, the first block of documents presented to a user is a set of 6 documents consisting of the 5 lowest ranking documents, plus the highest ranking document injected to maximize the difference between relevant and non-relevant.

- In stage *III*, each of the blocks are randomized independently, and then combined to form the final assessment list.

Experimental design summary. Figure 3 shows the experimental design procedure and a sample of the experiment interface. As can been seen, each assessor completes an introduction and training phase, followed by a pre-experiment questionnaire. Assessors use a unique user identifier assigned to them to complete the pre-experiment questionnaire, which gathers basic demographic and background information about assessor's web and search engine usage experience. Assessors are then directed to the first topic, followed by the post-hoc questionnaires described in Table 1. This is followed by the second topic, and another post-hoc questionnaire. A 10 minute break is provided between the two topics.

4 RESULTS AND DISCUSSION

Each assessor judged two of the eight topics, therefore we collected 5,760 data points (96 assessors × 2 topics × 30 documents). We focused on the two main research variables as effects of the document order (ILR, DLR and RLR) and difficulty (user difficulty: easy and hard, and system difficulty: easy, medium, and hard).

To investigate our research questions, we introduced a method to classify the user topic difficulty as user easy, user medium and user hard topics for RQ1, and we measured the inter-rater agreement

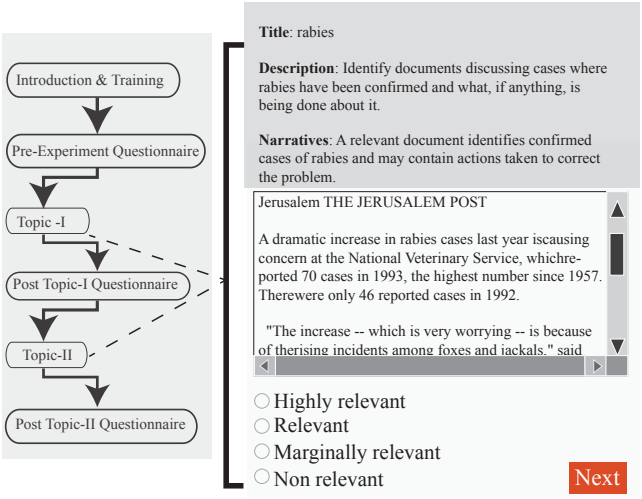

Figure 3: Procedure of the experimental design.

between our assessors for the RQ2 regarding the effect of presentation order. We then compared our user's judgments and the gold standard, for RQ3 – the interaction between the topic difficulty and document ordering.

Several analysis techniques are used in this study. First, to quantify the notion of user topic difficulty, we used structural equation modeling (*SEM*) [3, 21] with scaling and t-testing to compare the ratings of the lab and CrowdFlower assessors. Second, we adopted a generalized linear mixed model (*GLMM*) [6] with a binomial distribution and a logit function when comparing for gold standard agreement. We acknowledge that there may be individual differences which are caused by using between-subject design, and participant familiarity with web search. To consider the individual differences, we adopted a mixed model instead of a generalized linear model (GLM), because the random effect between subjects (σ_s^2) were greater than the standard error. If a significant effect was observed, we ran a post-hoc analysis test using standard error of differences (SEDs) to find the specific pairs responsible for the difference. Third, Krippendorff's α [20, 43] was used as a chance-corrected measure of agreement, between two or more assessors, and between assessors and the gold standard.

4.1 User Topic Difficulty

We conducted two crowdsourcing experiments using CrowdFlower to analyze topic difficulty from a user perspective. In the first experiment, assessors were provided with only TREC topics to rate topics using the questions in Table 1. We employed the same questions for the second experiment, but assessors were shown one highly relevant and one non-relevant document along with each of the topics. Ordering of the relevant and non-relevant documents was balanced, so that half of the participants received the documents in each ordering. The aim of including these example documents was to study whether this additional information has an impact on user perception of topic difficulty.

We computed the Cronbach's α score between Q4, Q5 and Q6 to measure the consistency of the questions for user topic difficulty, and the scores are 0.90 and 0.95 for the first and second experiments,

Table 4: Factor loadings (λ_i) and standard error (δ_i) of the model $D_i = \lambda_i \xi_i + \delta_i$

	experiment 1		experiment 2	
Qns.	λ_i	δ_i	λ_i	δ_i
$Q4$	1.000		1.000	
$Q5$	0.957	0.117	0.955	0.088
$Q6$	0.770	0.123	0.886	0.080

respectively. We confirmed consistency between user and system difficulty, and finally we applied *SEM* to specify a model for the composite factor – user topic difficulty. A Composite variable difficulty (D_i) is estimated using the model:

$$D_i = \lambda_i \xi_i + \delta_i$$

Here δ_i represents error associated to a rating, λ_i represents factor loadings – latent variable estimates of questions – and ξ_i is the scaling factor. Table 4 shows the factor loadings and standard error of the rating estimates of questions in the model. The scaling factors (ξ_i) for experiments one and two is 1.103 and 1.056 respectively; which is the sum of the number of questions (3) divided by the sum of the factor loadings in each experiment.

The averages across the two experiments were then used to sort topics by user difficulty. Lower values of the average score indicate that users find the topics easier than higher average values of the composite. The sorted list is divided into three equal segments where equal number of topics will be in each of the segments. The top 14 topics are put into the first segment and labeled as user easy (*UE*), the bottom 14 topics into the third segment and labeled as user hard (*UH*), and the remaining middle 13 topics in the second segment and labeled as user medium (*UM*). This method is analogous to the method by Carterette et al. [8] to assign category labels to topics using the AAP score (except we used a composite of user ratings).

Easy and *hard* topics are the main focus of our study. Since distinctions between system medium and user medium difficulty are often ambiguous, and less interesting for this study, two system medium topics were included in the user study for calibration purposes. Eight topics were ultimately selected from the different classes shown in Table 3. The difference between assessor ratings of topic difficulty between the CrowdFlower assessors and participants in the subsequent lab-based user study were analyzed using an unpaired t-test. The result ($t(13.481) = 0.8273, p = 0.422$) showed no significant difference between the self-ratings of the two assessor groups, for the eight topics chosen for the lab user study.

In response to RQ1, the evidence suggests that user difficulty is a shared construct between CrowdFlower and lab user study groups for the eight topics. In addition, the results reinforce the belief that questions Q4, Q5 and Q6 in Table 1 can be used to measure user topic difficulty.

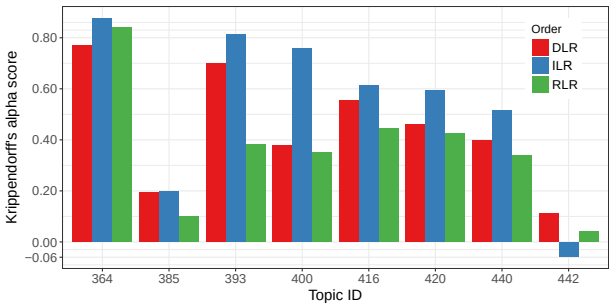

Figure 4: Average pairwise agreement between our assessors and TREC gold standard judgments, measured across an aggregate of topics and order using Krippendorff's α on ratings on a binary scale, flattening 0 and 1 to 0; and 2 and 3 to 1.

4.2 Document Presentation Order

This section further explores RQ2 – how does presentation ordering affect inter-rater agreement when judging the relevance of documents? Inter-rater agreement can be measured between our assessors and the TREC assessors; or between the assessors who completed the assessment tasks. We call the agreement between our assessors and TREC assessors *gold agreement* and the agreement between our assessors *inter-rater agreement*.

Gold agreement. Gold agreement is measured by comparing each assessor's judgment with an existing TREC judgment. The TREC judgments are binary, while our judgments are collected using a 4-level ordinal scale. The ordinal scales therefore need to be folded into binary; we combine non-relevant and marginally relevant documents into a single non-relevant category, while relevant and highly relevant documents are combined into a single relevant category [39]. Following the binary folding of ratings, Krippendorff's α is computed between each assessor and the gold ratings. Finally, the result is aggregated across the topics, broken by the presentation orders.

As can be seen in Figure 4, for all topics except 442, ILR ordering has the highest gold agreement. The mean gold agreement scores across the 8 topics are 0.539 for ILR, 0.446 for DLR, and 0.365 for RLR. Topic 442 is one of the system hard topics; and the overall pairwise agreement is the lowest compared to the other topics. In addition, comparing the orderings within the topic itself, the agreement is also the lowest (−0.057) in ILR as compared to DLR and RLR. This overall lowest gold agreement of topic 442 might be due to the number of relevant documents in the assessment list. According to the findings of Al-Maskari et al. [1], topics with fewer relevant documents result in proportionally more documents being judged differently when compared to TREC assessments. Topic 442 and 385 have the lowest number of relevant documents compared to the other topics in our experiment. Table 5 shows the count of relevant documents in the TREC QRELs, and in our assessment list as judged by the TREC assessors. Note that count of the number of relevant documents in our assessment list and TREC QRELs is performed after the the experiment is completed.

Table 5: Number of judged relevant by TREC assessors in the original TREC QREL files and in the sample 30 documents used for the lab user study as generated by the NTCIRPOOL and our document sampling method.

		Total docs. in			Relevant docs. in	
Topic ID	AAP	pool	uniq.	QREL (TREC)	QREL (TREC)	sample
#364	0.45	5029	961	1513	35	7
#420	0.38	5897	812	1136	33	4
#393	0.04	4977	1507	2291	71	6
#442	0.01	5959	2101	2679	94	1
#385	0.21	5100	921	1326	86	2
#400	0.42	5045	669	1009	125	9
#416	0.30	5890	1002	1235	42	5
#440	0.09	5964	1443	1830	54	4

Table 6: Number of assessors with a statistically significant pairwise difference compared to TREC assessors as measured using unpaired t-test.

	Topic ID					
Order	#385	#393	#400	#420	#442	Total
ILR	6	0	0	0	8	14
DLR	1	1	2	0	5	9
RLR	4	2	4	1	3	14
Total	11	3	6	1	16	37

From Table 5 and Figure 4 it can be seen that the agreement in ILR order is higher when there are 6 or more relevant documents in the assessment list.

Out of a total of 192 assessors (considering each topic-assessor combination as distinct), 37 assessors differed significantly in their assessment with TREC assessors, as shown in Table 6. Our assessors exhibit no significant difference on their relevance judgments when compared to TREC assessors for three of the topics (#364, #416, #420).

Inter-rater agreement. Inter-rater agreement is measured using Krippendorff's α and the results are shown in Table 7. The highest overall (All) inter-rater agreement is reported for ILR ordering (0.810), much higher than for either DLR (0.632) or RLR (0.430). In addition, between user easy and user hard topics, agreement is higher in easy topics than hard topics for the other two ordering (DLR and RLR). This is an interesting result, as ILR order might help maximize agreement in user hard topics; which is the behavior we would most like to address (decreasing relevance drift). The results in Table 7 also show that the agreement for DLR ordering is higher than in RLR ordering. Assessors using DLR ordering have the benefit of learning more about the topic in the first few documents, which may not be the case in RLR ordering. This might be the reason for higher agreement in DLR than in RLR. It is worth noting that this effect may not be true for real TREC assessors, as they are typically topic originators, and therefore may have a clearer notion of relevance for the topic before the assessment exercise begins. Our findings focus on the case of gathering relevance assessments for topics that do not originate from the assessor. These are very different scenarios.

Table 7: Inter-rater agreement measured between assessors using Krippendorff's alpha (α) across individual topics with ratings on a 4-level ordinal scale. Each topic is assessed by 8 assessors

Topic ID	Title	Document Ordering		
		ILR	DLR	RLR
#364	Rabies	0.965	0.843	0.883
#420	carbon monoxide poisoning	0.755	0.533	0.460
#393	mercy killing	0.836	0.731	0.360
#442	heroic acts	0.727	0.596	0.225
#385	hybrid fuel cars	0.591	0.511	0.323
#400	amazon rain forest	0.866	0.510	0.315
#416	three gorges project	0.941	0.679	0.475
#440	child labor	0.802	0.652	0.449
UE	User Easy	0.775	0.643	0.450
UH	User Hard	0.869	0.614	0.413
SE	System Easy	0.862	0.629	0.553
SM	System Medium	0.766	0.595	0.399
SH	System Hard	0.788	0.659	0.345
All		0.810	0.632	0.430

In response to RQ2, gold agreement is higher for easy topics than hard (user or system), and ILR order has the highest overall gold agreement (pairwise mean = 0.539) compared to DLR (pairwise mean = 0.446) and RLR (pairwise mean = 0.365). In addition, the overall results show that inter-rater agreement was highest in ILR. In other words, presentation ordering using ILR helps assessors converge on the notion of relevance more quickly when compared to the other two orderings.

4.3 Topic Difficulty and Document Ordering Interaction

To answer RQ3, which focuses on the interaction between topic difficulty and presentation order, we adopted a GLMM (full model) for the gold agreement comparisons rather than using inter-rater agreement, in order to consider three research variables simultaneously, i.e., the order, user difficulty, system difficulty and their interactions.

Before exploring the interactions, the order and system difficulty effects were compared with the gold assessments, as shown in Table 8 ($\sigma_s^2 = 0.056$, $X^2 = 3.62$, df = 2, $p < 0.01$, and $X^2 = 11.14$, df = 2, $p < 0.001$, respectively). The results on order effects indicate that users judging in ILR order exhibited higher gold agreement than when using RLR order (86.9% vs 83.3% for the ILR and RLR

Table 8: Gold agreement for each order, broken by user and system difficulties, and the interactions between the order and difficulty.

		ILR	DLR	RLR	p-value		
					Order	Difficulty	Interaction
User difficulty [%]	UE	85.92	87.92	84.83	*	0.259	**
	UH	88.60	83.89	80.80			
System difficulty [%]	SE	92.10	86.50	85.69	*	***	**
	SM	81.67	85.83	78.33			
	SH	85.28	86.67	84.31			

*Significant at 0.05 level. ** Significant at 0.01 level. *** Significant at 0.001 level.

orders, respectively), although there is no significant difference between ILR and DLR orderings (86.4%). Surprisingly, the assessors showed similar gold agreement with system easy and hard topics, but the agreement with system medium topics was lower than the others (88.1%, 85.4%, and 81.9% for system easy, hard, and medium, respectively).

For the interaction between the order and user difficulty, we found significant effects relative to the gold assessments ($X^2 = 5.30$, df = 2, $p < 0.01$). As shown in Table 8, RLR order displays only 84.83% for user easy topics, which is significantly lower than DLR order (87.92%). For user hard topics, assessors showed the highest gold agreement (88.60%) with the ILR order, whereas they recorded 83.89% and 80.80% with the DLR and RLR orders, respectively. This indicates that the ILR order is superior when assessing user hard topics.

Order and system difficulty also show significant interactions with respect to gold agreement ($X^2 = 4.17$, df = 4, $p < 0.01$). Table 8 shows that the assessors, for system easy topics, had higher gold agreement with ILR order (92%) than with other orders (86% for both DLR and RLR orders). Our participants also recorded the worst gold agreement with the RLR order (78%) in the cases of using system medium topics, and they exhibit no significant difference with system hard topics.

Summarizing the above results for RQ3, assessors using ILR tended to exhibit significantly higher gold agreement for user hard and system easy topics, and performed better or similarly for other levels of topic difficulty. In addition, RLR order seems to be the worst with respect to gold agreement across all combinations of topic difficulty.

For the effect of presentation order, we cannot say that ILR order is overwhelmingly superior when considering each result by user and system difficulties as shown in the result of inter-rater agreement, but ILR order is generally better than RLR order, and is better or similar to the DLR order in the gold agreement. In particular, hard topics seem to be much more consistent when using ILR, which is an important finding.

5 CONCLUSIONS AND LIMITATIONS

This study was designed to investigate three key research questions about defining and measuring user topic difficulty, and the effects of document presentation order and their interactions.

Limitations. We carefully designed two large scale user studies; however, a number of limitations should be considered. First, we acknowledge that these results may not represent the search behavior of the general public, although we conducted the crowdsourcing and laboratory user studies with large numbers of participants (103 and 96, respectively). Second, we assumed that participants can maintain their concentration for an hour in a lab-based setting, although to mitigate this issue participants were able to take a break between each topic exercise, and they were allowed to leave at any point during the experiment if they felt uncomfortable. Third, while we carefully selected topics to represent different levels of user and system difficulty, in order to meet available resources the final study used a selection of eight topics, and so the presence of other topic effects cannot be ruled out.

Conclusions. For the first research question, we analyzed topic difficulty from the perspective of users through a questionnaire. Our findings across both crowdworker and lab-based participants suggest that user difficulty is a shared construct which can be measured using related variables such as perceived difficulty to search information, difficulty to determine document relevance, and an overall perception of topic difficulty given a topic. Our results show that the chosen representations of user and system difficulty are orthogonal; for our study we therefore carefully selected a range of topics that cover a range of both user and system difficulty.

For the second research question regarding the effect of presentation ordering, we explored using gold agreement (agreements between our assessors and the TREC judgments) and inter-rater agreement (the agreement between the assessors' judgments). The key finding is that our proposed presentation order (ILR) is the most effective order to maximize agreement (both inter-rater and agreement with TREC judgments) and consistency of relevance judgments across the different classes of topic difficulty investigated in this research. In particular, a level of inter-rater agreement of 0.81 was obtained using our new method, compared to 0.63 and 0.42 for decreasing relevance and random orderings. This finding and has direct application for the design of reusable test collections.

For the third research question about the interaction between presentation order and topic difficulty, we investigated the effect of interaction among document presentation ordering, user difficulty and system difficulty. We found significant interactions between order and topic difficulty (user and system), that is, the proposed

ordering, ILR, contributes to the higher consistency of relevance judgments than the other two orderings, especially for both system and user hard topics.

Overall, our proposed method of document ordering improves the consistency of relevance judgments and agreement between two or more assessors. The performance of ILR is superior compared to DLR and RLR when comparing the judgments with the gold standard, and the technique also leads to higher inter-rater agreement. The technique therefore offers a direct benefit when creating new test collections, without requiring any additional resources.

Acknowledgements. This work was supported by the Australian Research Council's *Discovery Projects* Scheme (DP170102231, DP170102726 and DP180102687), and by a grant from the Mozilla Foundation.

REFERENCES

[1] A. Al-Maskari, M. Sanderson, and P. Clough. 2008. Relevance judgments between TREC and Non-TREC assessors. In *Proc. SIGIR*. 683–684.
[2] O. Alonso and R. Baeza-Yates. 2011. Design and implementation of relevance assessments using crowdsourcing. In *Proc. ECIR*. 153–164.
[3] J. C. Anderson and D. W. Gerbing. 1988. Structural equation modeling in practice: A review and recommended two-step approach. *Psy. Bulletin* 103, 3 (1988), 411.
[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. 2008. Relevance assessment: are judges exchangeable and does it matter. In *Proc. SIGIR*. 667–674.
[5] M. Braschler. 2003. CLEF 2003–Overview of results. In *Workshop of Cross-Lang. Eval. Forum on EU. Lang*. Springer, 44–63.
[6] N. E. Breslow and D. G. Clayton. 1993. Approximate inference in generalized linear mixed models. *JASA* 88, 421 (1993), 9–25.
[7] B. Carterette, P. N. Bennett, D. M. Chickering, and S. T. Dumais. 2008. Here or there. In *Proc. ECIR*. 16–27.
[8] B. Carterette, V. Pavlu, H. Fang, and E. Kanoulas. 2009. Million Query Track 2009 Overview.. In *Proc. TREC*.
[9] C. Cleverdon. 1967. The Cranfield tests on index language devices. In *Aslib proceedings*, Vol. 19. MCB UP Ltd, 173–194.
[10] A. Crescenzi, D. Kelly, and L. Azzopardi. 2016. Impacts of time constraints and system delays on user experience. In *Proc. CHIIR*. 141–150.
[11] J. S. Culpepper, S. Mizzaro, M. Sanderson, and F. Scholer. 2014. TREC: Topic engineeRing ExerCise. In *Proc. SIGIR*. 1147–1150.
[12] T. T. Damessie, T.P. Nghiem, F. Scholer, and J. S. Culpepper. 2017. Gauging the Quality of Relevance Assessments using Inter-Rater Agreement. In *Proc. SIGIR*. 1089–1092.
[13] T. T. Damessie, F. Scholer, , and J. S. Culpepper. 2016. The Influence of Topic Difficulty, Relevance Level, and Document Ordering on Relevance Judging. In *Proc. ADCS*. 41–48.
[14] T. T. Damessie, F. Scholer, K. Järvelin, and J. S. Culpepper. 2016. The effect of document order and topic difficulty on assessor agreement. In *Proc. ICTIR*. 73–76.
[15] M. Eisenberg and C. Barry. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. *J. Am. Soc. Inf. Sci.* 39, 5 (1988), 293–300.
[16] N. Gisev, J. S. Bell, and T. F. Chen. 2013. Interrater agreement and interrater reliability: Key concepts, approaches, and applications. *Res. in Soc. and Admin. Phar.* 9, 3 (2013), 330 – 338. DOI:http://dx.doi.org/https://doi.org/10.1016/j.sapharm.2012.04.004
[17] C. Grady and M. Lease. 2010. Crowdsourcing document relevance assessment with mechanical turk. In *Proc. CSLDAMT*. Association for Computational Linguistics, 172–179.
[18] D. Harman. 2011. Information retrieval evaluation. *Synthesis Lectures on Information Concepts, Retrieval, and Services* 3, 2 (2011), 1–119.
[19] D. Hawking. 2000. Overview of the TREC-9 Web Track.. In *Proc. TREC*.
[20] A. F. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Comm. Methods and Measures* 1, 1 (2007), 77–89.
[21] J. J. Hox and T. M. Bechger. 2007. An introduction to structural equation modeling. (2007).
[22] M. Huang and H. Wang. 2004. The influence of document presentation order and number of documents judged on users' judgments of relevance. *J. Am. Soc. Inf. Sci. and Tech.* 55, 11 (2004), 970–979.
[23] N. Kando, K. Kuriyama, T. Nozue, K. Eguchi, H. Kato, and S. Hidaka. 1999. Overview of IR tasks at the first NTCIR workshop. In *Proc. NTCIR*. 11–44.
[24] E. Kanoulas, D. Li, L. Azzopardi, and R. Spijker. 2017. CLEF 2017 technologically assisted reviews in empirical medicine overview. In *CEUR Workshop Proceedings*, Vol. 1866. 1–29.
[25] R. V. Katter. 1968. The influence of scale form on relevance judgments. *Information Storage and Retrieval* 4, 1 (1968), 1–11.
[26] G. Kazai. 2011. In search of quality in crowdsourcing for search engine evaluation. In *Proc. ECIR*. 165–176.
[27] J. Kekäläinen. 2005. Binary and graded relevance in IR evaluationsfi!?comparison of the effects on ranking of IR systems. *Inf. Proc. & Man.* 41, 5 (2005), 1019–1033.
[28] D. Kelly. 2009. Methods for evaluating interactive information retrieval systems with users. *Found. Trends in Inf. Ret.* 3, 1–2 (2009), 1–224.
[29] J. Kottner, L. Audigé, S. Brorson, A. Donner, B. J. Gajewski, A. Hróbjartsson, C. Roberts, M. Shoukri, and D. L. Streiner. 2011. Guidelines for reporting reliability and agreement studies (GRRAS) were proposed. *Int. J. of Nurs. Stud.* 48, 6 (2011), 661–671.
[30] S. Mizzaro and S. Robertson. 2007. Hits hits TREC: exploring IR evaluation results with network analysis. In *Proc. SIGIR*. 479–486.
[31] A. M. Rees and D. G. Schultz. 1967. A Field Experimental Approach to the Study of Relevance Assessments in Relation to Document Searching. *Final Report to the NSF* I (1967).
[32] S. Robertson. 2008. On the history of evaluation in IR. *J. Inf. Sci.* 34, 4 (2008), 439–456.
[33] R. S. Roy, M. Choudhury, P. Majumder, and K. Agarwal. 2013. Overview of the fire 2013 track on transliterated search. In *Proc. FIRE*. 4.
[34] T. Sakai and N. Kando. 2008. Are Popular Documents More Likely To Be Relevant? A Dive into the ACLIA IR4QA Pools. In *Proc. NTCIR*.
[35] T. Sakai and C. Lin. 2010. Ranking Retrieval Systems without Relevance Assessments: Revisited. In *Proc. EVIA*. 25–33.
[36] M. Sanderson. 2010. *Test collection based evaluation of information retrieval systems*.
[37] T. Saracevic. 2007. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *J. Am. Soc. Inf. Sci. and Tech.* 58, 13 (2007), 1915–1933.
[38] F. Scholer, D. Kelly, W. Wu, H. S. Lee, and W. Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In *Proc. SIGIR*. 623–632.
[39] F. Scholer and A. Turpin. 2009. Metric and relevance mismatch in retrieval evaluation. *Proc. AIRS* (2009), 50–62.
[40] M. D. Smucker, G. Kazai, and M. Lease. 2012. *Overview of the trec 2012 crowdsourcing track*. Technical Report. TEXAS UNIV AT AUSTIN SCHOOL OF INFORMATION.
[41] E. Sormunen. 2002. Liberal relevance criteria of TREC-: Counting on negligible documents?. In *Proc. SIGIR*. 324–330.
[42] M. Stefano. 1997. Relevance: The whole history. *J. Am. Soc. Inf. Sci.* 48, 9 (1997), 810–832.
[43] K. D. Swert. 2012. Calculating inter-coder reliability in media content analysis using Krippendorff's Alpha. *Center for Politics and Communication* (2012).
[44] M. Tavakol and R. Dennick. 2011. Making sense of Cronbach's alpha. *Int. J. of Med. Edu.* 2 (2011), 53.
[45] A. Taylor. 2012. User relevance criteria choices and the information search process. *Inf. Proc. & Man.* 48, 1 (2012), 136–153.
[46] P. Vakkari and E. Sormunen. 2004. The influence of relevance levels on the effectiveness of interactive information retrieval. *J. Am. Soc. Inf. Sci. and Tech.* 55, 11 (2004), 963–969.
[47] E. M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Inf. Proc. & Man.* 36, 5 (2000), 697–716.
[48] E. M. Voorhees and D.K. Harman (Eds.). 2005. *TREC: Experiment and evaluation in information retrieval*. Vol. 1. MIT press Cambridge.