On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation

EDDY MADDALENA, University of Udine, Italy
STEFANO MIZZARO, University of Udine, Italy
FALK SCHOLER, RMIT University, Australia
ANDREW TURPIN, University of Melbourne, Australia

Magnitude estimation is a psychophysical scaling technique for the measurement of sensation, where observers assign numbers to stimuli in response to their perceived intensity. We investigate the use of magnitude estimation for judging the relevance of documents for information retrieval evaluation, carrying out a large-scale user study across 18 TREC topics and collecting over 50,000 magnitude estimation judgments using crowdsourcing. Our analysis shows that magnitude estimation judgments can be reliably collected using crowdsourcing, are competitive in terms of assessor cost, and are, on average, rank-aligned with ordinal judgments made by expert relevance assessors.

We explore the application of magnitude estimation for IR evaluation, calibrating two gain-based effectiveness metrics, nDCG and ERR, directly from user-reported perceptions of relevance. A comparison of TREC system effectiveness rankings based on binary, ordinal, and magnitude estimation relevance shows substantial variation; in particular, the top systems ranked using magnitude estimation and ordinal judgments differ substantially. Analysis of the magnitude estimation scores shows that this effect is due in part to varying perceptions of relevance: different users have different perceptions of the impact of relative differences in document relevance. These results have direct implications for IR evaluation, suggesting that current assumptions about a single view of relevance being sufficient to represent a population of users are unlikely to hold.

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

General Terms: Measurement, performance, experimentation

Additional Key Words and Phrases: Magnitude estimation, evaluation, relevance assessments, relevance

ACM Reference Format:
... *ACM Trans. Inf. Syst.* V, N, Article 13 (Month 2016), 30 pages.
DOI: http://dx.doi.org/10.1145/3002172

1. INTRODUCTION

Relevance is an important concept in information retrieval (IR), and relevance judgments form the backbone of test collections, the most widely-used approach for the evaluation of IR system effectiveness. Document relevance judgments are typically made using ordinal scales, historically at a binary level, and more recently with multiple levels [Voorhees and Harman 2005]. However, despite its importance, operationalizing relevance for the evaluation of IR systems remains a complicated issue; for example, when using an ordinal scale it is unclear how many relevance categories should be chosen [Tang et al. 1999].

Magnitude estimation is a psychophysical technique for measuring the sensation of a stimulus. Observers assign numbers to a series of stimuli, such that the numbers reflect the perceived difference in intensity of each item. The outcome is a ratio scale of the subjective perception of the

---

This paper is a revised and extended version of a SIGIR2015 paper [Turpin et al. 2015]. Author's addresses: Eddy Maddalena, University of Udine, Udine 33100, Italy, eddy.maddalena@uniud.it; Stefano Mizzaro, University of Udine, Udine 33100, Italy, mizzaro@uniud.it; Falk Scholer. RMIT University, Melbourne, Victoria 3001, Australia, falk.scholer@rmit.edu.au; Andrew Turpin, University of Melbourne, Melbourne, Victoria 3010, Australia, aturpin@unimelb.edu.au.

DOI: http://dx.doi.org/10.1145/0000000.0000000

stimulus; if a magnitude of 50 is assigned to one stimulus, and 10 to another, then it can be concluded that the two items are perceived in a 5:1 ratio of sensation [Moskowitz 1977]. While initially developed for the measurement of the sensation of physical stimuli such as the intensity of a light source, magnitude estimation has been successfully applied to the measurement of non-physical stimuli including the usability of interfaces [McGee 2003].

Being able to derive ratio scales of subjective perceptions of the intensity of stimuli, magnitude estimation may be a useful tool for measuring and better understanding document relevance judgments. The application of magnitude estimation in the IR field has been limited to the consideration of the relevance of carefully curated abstracts returned from bibliographic databases [Eisenberg 1988], and our own small-scale pilot study [Maddalena et al. 2015; Scholer et al. 2014]. While suggestive, these studies have been limited in terms of demonstrating the broader utility of the approach, or its direct application to IR evaluation. In this work, we investigate the larger-scale application of magnitude estimation to document relevance scaling, reporting on a user study over 18 TREC topics and obtaining judgments for 4,269 documents. This is also the first work to consider the direct application of magnitude estimation to the evaluation of IR systems, and to rely on crowdsourcing workers to gather a large amount of magnitude estimation relevance assessments. Specifically, we consider four research questions.

RQ1. Is the magnitude estimation technique suitable for gathering document-level relevance judgments, and are the resulting relevance scales reasonable with respect to our current knowledge of relevance judgments?
RQ2. Is crowdsourcing a viable method to gather robust magnitude estimation values? Crowdsourcing might exacerbate some of the potential complications of using magnitude estimation, such as the subjective scale used by each judge, or the difficulty in using a ratio scale without proper training. Do the workers agree with previous judgments, and are redundant workers required to obtain agreement with previous judging methods?
RQ3. How does IR system evaluation change when ratio scale magnitude estimation relevance judgments are used to calibrate the gain levels of the widely-used nDCG and ERR evaluation metrics, compared to using arbitrarily set gain values for a pre-chosen number of ordinal levels?
RQ4. Can magnitude estimation relevance scores provide additional insight into user perceptions of relevance, into actual gain values, and into individual gain profiles?

We also investigate the qualitative comments that users of the magnitude estimation relevance judging approach were required to make to justify their scores, and present an analysis of the relationship between comments, scores and source documents.

In the next section, background material and related work are presented. Details of our user study and experimental methodology are provided in Section 3, and the first descriptive results are presented in Section 4. The full analysis and discussion of our results, corresponding to the four research questions, are detailed in Sections 5 to 8. Our conclusions and directions for future work are included in the final section of the paper.

2. BACKGROUND

2.1. Relevance

Relevance is an important concept in information science and information retrieval [Saracevic 2007]. IR system effectiveness is most often measured with reference to a test collection, consisting of a set of search topics, a static set of documents over which to search, and human-generated assessments that indicate, for an answer document returned by a search system in response to a topic, whether the document is relevant [Voorhees and Harman 2005]. Typically, these assessments are made based on "topical" relevance, that is, a judgment of whether the document contains any information that is "about" the material that the search topic is asking for. Other aspects of relevance, such as novelty or contextual factors, are not considered [Saracevic 2007]. Based on the relevance

judgments, each ranked answer list returned by a system is aggregated into a number that reflects the system's performance, using a chosen effectiveness metric. The metrics themselves instantiate either implicit or explicit assumptions about searcher behavior, for example that a relevant document that is returned by a system lower down a ranked list is of less value than if the same document had been returned earlier.

Historically, relevance judgments were often made using a *binary* scale, where a document is classed as either being relevant to the search topic or not, and many effectiveness metrics use this notion of relevance, including *precision*, *recall*, *mean average precision* (MAP), and *precision at a specified cutoff* [Voorhees and Harman 2005]. More recently, based on the observation that searchers can generally distinguish between more than two levels of relevance, evaluation metrics that incorporate *multi-level* relevance have been proposed [Järvelin and Kekäläinen 2002]. Here, relevance is typically measured on an ordinal scale; metrics that include this more fine-grained notion of relevance include *normalized discounted cumulative gain* (nDCG) [Järvelin and Kekäläinen 2002], *expected reciprocal rank* (ERR) [Chapelle et al. 2009], and the *Q-measure* [Sakai 2007]. Some examples of previous choices for the number of levels in ordinal relevance scales include 3 (TREC Terabyte Track [Clarke et al. 2005]), 4 (TREC newswire re-assessments by Sormunen [2002]), and 6 (TREC Web Track [Collins-Thompson et al. 2014]). Tang et al. [1999] studied the self-reported confidence of psychology students when assessing the relevance of bibliographic records on ordinal scales from 2 to 11 levels, and observed that for this specific task, confidence is maximized with a 7-point scale. However, the issue is far from settled: in a broader survey of the optimal number of levels in an ordinal response scale, Cox [1980] concludes that there is no single number of response alternatives that is appropriate for all situations.

2.2. Magnitude Estimation

Magnitude estimation (ME) is a psychophysical technique for the construction of measurement scales for the intensity of sensations. An observer is asked to assign numbers to a series of presented stimuli. The first number can be any value that seems appropriate, with successive numbers then being assigned so that their relative differences reflect the observer's subjective perceptions of differences in stimulus intensity [Ehrenstein and Ehrenstein 1999]. A key advantage of using magnitude estimation is that the responses are on a ratio scale of measurement [Gescheider 1997], meaning that all mathematical operations may be applied to such data, and parametric statistical analysis can be carried out. In contrast, for ordinal (or ranked category) scales certain operations are not defined; for example, the median can be used as a measure of central tendency for ordinal data, but the mean is not meaningful since the distance between the ranked categories is not defined [Sheskin 2007].

Proposed by Stanley Stevens in the 1950s, magnitude estimation has a long history, and is the most widely-used psychophysical ratio scaling technique [Gescheider 1997]. Initially developed to measure perceptions of physical stimuli, such as the brightness of a light or the loudness of a sound, magnitude estimation has also been successfully applied to a wide range of non-physical stimuli in the social sciences (including occupational preferences, political attitudes, the pleasantness of odors, and the appropriateness of punishments for crimes [Stevens 1966]), in medical applications (such as levels of pain, severity of mental disorders, and emotional stress from life events [Gescheider 1997]), in user experience research (for example, as a measure of usability in HCI [McGee 2003], and for health-care applications [Johnson and Bojko 2010]), and in linguistics (including judging the grammaticality of sentences [Bard et al. 1996]).

In information retrieval research, Eisenberg [1988] investigated magnitude estimation in the context of judging the relevance of document citations from a library database (including fields such as author, title, keywords and abstract), and concluded that participants are able to effectively use magnitude estimation in such a scenario. A related technique was used by Spink and Greisdorf [2001], where participants in a user study were required to fill in a worksheet with information about the relevance of resources that were retrieved from a library database for personal research projects; this included indicating the level of relevance on a 4-level ordinal scale, providing feedback about other

levels of relevance such as utility and motivation, and marking the level of relevance on a 77mm line. The line was then decoded into numbers at a 1mm resolution so was in effect a 78-level ordinal scale.

To the best of our knowledge, the only previous investigation of magnitude estimation for document-level relevance judging in the context of IR evaluation was our own small pilot study [Maddalena et al. 2015; Scholer et al. 2014]. That study indicated that magnitude estimation could be useful for measuring document-level relevance, but used only 3 information need statements and 33 documents. In this paper we carry out a much larger-scale study, across 18 TREC topics and 4269 documents. Unlike previous studies, our work also explores the application of this scaling approach to the direct calibration of effectiveness metrics for IR system evaluation, and the relationship between the magnitude relevance space and the concept of gain.

This paper is a revised and extended version of research that appeared at SIGIR 2015 [Turpin et al. 2015]. In particular, this version adds: a new research question of whether crowdsourcing is a viable method to gather robust magnitude estimation values (RQ2); related background on crowdsourcing; more information on the ME score distributions reported in Section 4, including Figures 1 and 2 and their discussion; a new Section 6 including Figure 5 and re-computation of Figure 6; new material in Section 7, including re-computation of Figures 10 and 11; a revised second half of Section 8; an Appendix giving exact participant instructions; and additional references.

2.3. Crowdsourcing

The third ingredient of our work is crowdsourcing. namely the outsourcing of a task usually performed by experts to a crowd by means of an open call. Although criticized by some [Keen 2008], crowdsourcing has been shown to be effective for several tasks, at least under some conditions, and it has been (and still is) much used as a method to collect relevance judgments.

For example Alonso and Mizzaro [2012] compared relevance judgments gathered by the crowd vs. more experienced TREC assessors, finding comparable accuracy. Others have attempted to measure relevance in various ways, including graded judgments [McCreadie et al. 2011], preference-based judgments [Anderton et al. 2013], and multidimensional ones [Zhang et al. 2014]. Collecting relevance labels through crowdsourcing has also been used in practice, for example in the TREC blog track [McCreadie et al. 2011], or in the judgment task of the TREC Crowdsourcing Track [Smucker et al. 2014]. Still in the IR field, it has been shown that crowdsourcing can be effective not only to collect relevance judgments, but also for more complex tasks [Zuccon et al. 2013].

Of course, quality is an issue: workers might be too inexperienced for some tasks, or even behave maliciously to finish the task without any accuracy, perhaps even in a random way, just to gain money. Several papers address this problem. For example, Clough et al. [2013] show that even if workers disagree, system rankings might not change. Kazai et al. [2013] study how several factors (e.g., pay, effort, worker qualifications, motivations, and interest) can affect the quality of the relevance judgments collected.

Appropriate aggregation methods can be used as countermeasures to low quality work. Hosseini et al. [2012] and Jung and Lease [2011] compare the classical aggregation method of majority voting with more sophisticated approaches that take into account a worker's expertise (based on expectation maximization and outlier detection, respectively), finding the latter to be more effective.

However, it is important to note that disagreement does not necessarily imply low worker quality. As pointed out for example by Aroyo and Welty [2015], disagreement might depend on the false assumption of a "unique truth", it might carry useful information that can be exploited, and attempts to eliminate disagreement, for example by detailed instructions, might not improve quality and simply lead to information loss.

In this work, we use crowdsourcing to gather magnitude estimation relevance assessments at scale. Apart from our own pilot studies [Maddalena et al. 2015; Scholer et al. 2014] we are not aware of any studies that uses crowdsourcing to gather magnitudes, let alone using crowdsourcing to gather relevance judgments expressed by magnitude estimation. Since magnitude estimation is a

Table I. Topics, number of documents, number of units, $\mathtt{H}_k$, $\mathtt{N}_k$, number of back buttons usage, number of failures in the (c) and (d) checks, as detailed in the text.

Topic	No. docs	No. units	$\mathtt{H}_k$	$\mathtt{N}_k$	Back			Fail		
					0	1	≥ 2	0	1	≥ 2
402	278	460	LA111689-0162	FBIS3-10954	452	8	0	426	18	16
403	111	182	LA092890-0067	LA071290-0133	178	4	0	120	19	43
405	214	354	LA061490-0072	FBIS3-13680	347	6	1	322	20	12
407	212	350	FT921-6003	FR940407-2-00084	346	3	1	324	9	17
408	188	310	FT923-6110	LA062290-0070	306	4	0	266	16	28
410	212	350	FBIS4-64577	FBIS4-44440	346	4	0	326	10	14
415	179	295	FBIS3-60025	FBIS4-10862	289	4	2	255	23	17
416	174	287	FBIS4-49091	LA112590-0107	286	1	0	267	12	8
418	243	402	LA102189-0167	FT924-6324	394	8	0	372	15	15
420	164	270	LA121590-0108	LA112690-0001	266	4	0	256	4	10
421	342	567	FT941-428	LA073189-0033	554	11	2	537	16	14
427	195	322	FT943-5736	LA080590-0077	315	7	0	214	52	56
428	253	419	FT943-9226	FBIS3-20994	412	6	1	396	13	10
431	203	335	FBIS3-46247	FT944-5962	333	2	0	296	27	12
440	264	437	FT942-3471	LA020589-0074	427	10	0	397	19	21
442	408	677	LA011390-0057	FT923-4524	672	5	0	629	25	23
445	210	347	FT924-8156	LA031989-0092	334	11	2	334	10	3
448	419	695	LA080190-0139	FBIS3-16837	687	8	0	652	21	22
Sum:	4269	7059			6944	106	9	6389	329	341
Percentage:					98	2	0	90	5	5

process that is unlikely to be familiar to the average crowd worker, analyzing the suitability of the crowd for completing such tasks is novel.

3. EXPERIMENTAL METHODOLOGY

In this section we describe the details of the experiments that were carried out to crowdsource the required relevance judgments using magnitude estimation.

3.1. Topics and Documents

Document-level relevance assessments have traditionally been made using binary or multi-level ordinal scales. To compare magnitude estimation judgments to these approaches, we selected a set of search tasks and documents from the TREC-8 ad hoc collection, which studies informational search over newswire documents. The TREC-8 collection includes binary relevance judgments made by NIST assessors [Voorhees and Harman 1999]. Subsequently, Sormunen [2002] carried out a re-judging exercise where a subset of topics and documents were judged by a group of six Master's students of information studies, fluent in English although not native speakers, on a 4-level ordinal relevance scale: N–not relevant (0); M–marginally relevant (1); R–relevant (2); H–highly relevant (3).

The 18 *search topics* used in our study were the subset of TREC-8 topics which also have Sormunen ordinal judgments available. They are listed in the first column of Table I. The *documents* for which we collected magnitude estimation judgments were the set of the top-10 items returned by systems that participated in TREC-8. This gives a total of 4,269 topic-document pairs, of which 3881 have binary TREC relevance judgments available, and 805 of which have Sormunen ordinal judgments available. The number of documents for each topic is shown in Table I (2nd column).

3.2. User Study

We carried out a user study through the CrowdFlower crowdsourcing platform during December 2014 and January 2015. The experimental design was reviewed and approved by the RMIT University ethics review board. Participants were paid $0.2 for each task *unit*, defined as a group of magnitude estimation judgments for 8 documents in relation to one topic. The number of units for each topic is shown in Table I (3rd column).

3.2.1. Practice task. After agreeing to take part in the study, a participant was shown a first set of task instructions, including a brief introduction to the experiment and explanation of the magnitude estimation process. Since magnitude estimation may not be familiar to participants, they were first asked to complete a practice task, making magnitude estimations of three lines of different lengths, shown one at a time. If they successfully completed this practice task (success was defined as assigning magnitudes such that the numbers were in ascending order when the lines were sorted from shortest to longest), participants were able to move on to the main part of the experiment. The full instructions shown to participants for the practice task are included in Appendix A.

3.2.2. Main task. The main task of the experiment required making magnitude estimations of the relevance of documents. Participants were informed, by means of a second set of instructions, that they would be shown an information need statement, and then a sequence of eight documents that had been returned by a search system in response to the information need statement, presented in an arbitrary order, and that their task was to *indicate how relevant these documents appear* in relation to the information need.

For the main task, the title, description and narrative of a TREC topic were first displayed at the top of the screen. After reading the information need, participants had to respond to a 4-way multiple-choice question, to test their understanding of the topic. The test questions focused on the main information concepts presented in the topic statements, and were intended to check that participants were engaging with the task, and as a mechanism to remove spammers. Participants who were unable to answer the test question correctly were unable to continue with the task.

Next, participants were presented with eight documents, one at a time. For each document, the participant was required to enter a magnitude estimation number in a text box displayed directly below the document, and a brief justification of why they entered their number into a larger text field. They were then able to proceed to the next document. A back button was available in the interface, with participants being advised that this should only be used if they wish to correct a mistake; under 3% of submitted jobs included use of this feature (details on the occurrences of clicks on the back button are shown in Table I, where the number of units with zero, one, or two or more back button clicks are shown).

After entering responses for eight documents, the task was complete. Participants were able to complete further tasks for other topics, up to a maximum of 18 tasks, as they were not able to re-assess the same topic.

3.2.3. Magnitude estimation assignments. Participants were instructed to assign ME scores as follows. *You may use any numbers that seem appropriate to you – whole numbers, fractions, or decimals. However, you may not use negative numbers, or zero. Don't worry about running out of numbers – there will always be a larger number than the largest you use, and a smaller number than the smallest you use. Try to judge each document in relation to the previous one. For example, if the current document seems half as relevant as the previous one, then assign a score that is half of your previously assigned score.* While some applications of magnitude estimation use a fixed modulus (specifying a particular number that is to be assigned to the first stimulus that is presented), this has been found to promote clustering of responses and the potential over-representation of some numbers [Moskowitz 1977]; we therefore allowed participants to freely choose their own values. The complete instructions that were shown to participants for both the practice and main tasks are included in Appendix A.

3.2.4. Document ordering. As explained above, each participant task *unit* required the judging of a set of eight documents for a particular search topic statement. The experiments used a randomized design, with documents presented in random order, to avoid potential ordering effects and first sample bias [Moskowitz 1977]. The document sets were constructed such that two of the documents in each set were a known ordinal H and N document for the topic; these documents (henceforth $\mathtt{H}_k$ and $\mathtt{N}_k$) were the same for all the participants working on the same topic (the TREC document identifiers for each topic are shown in Table I). This was to ensure that each participant saw at least

one document from the high and low ends of the relevance continuum. Ensuring that stimuli of different intensity levels are included in a task has been found to be important for the magnitude estimation process [Gescheider 1997], and also has an impact on the score normalization process, described below. Moreover, including the two $\mathtt{N}_k$ and $\mathtt{H}_k$ documents with "known" ordinal relevance values enabled a further data collection quality control check: after judging all eight documents, participants who had assigned magnitude estimation scores for the $\mathtt{N}_k$ document that were larger than for the $\mathtt{H}_k$ document were not able to complete the task. In total, at least 10 ME scores were gathered for each topic-document pair.

3.2.5. Quality checks. In total, four quality checks were included in the data gathering phase of our experiments:

(a) a practice task requiring magnitude estimation of line lengths;
(b) a multiple-choice test question to check a participant's understanding of the topic;
(c) a check that the magnitude estimations score for $\mathtt{H}_k$ was greater than that assigned to $\mathtt{N}_k$; and
(d) each participant had to spend at least 20 seconds on each of the 8 documents.

If any of (a) or (b) was unsuccessful, the participant could not continue with the task. They were allowed to restart from scratch, on a different unit and therefore on a randomly selected topic, if willing to do so; the same quality checks were applied again. If checks (c) or (d) were unsuccessful, the participant received the message "*Your job is not accurate enough. You can revise your work to finish the task*", and was allowed to use the back button and revise their previously assigned scores if they wished (the same two checks (c) and (d) were performed again in such a case); however, participants were not made aware which documents or scores were the "offending" responses. Less than 10% of units resulted in conditions (c) or (d) being triggered (see details in Table I, where the number of eventually successful units with 0, 1, or two or more failed checks are shown), and as already mentioned the back button was used in less than 3% of units. Finally, there was a syntax check that the numeric scores input by the participants were in the $(0, +\infty)$ range. If any of the checks were not successfully completed, no data for that unit was retained. Based on these extensive quality checks, we did not carry out any further filtering once the data had been collected.

4. GENERAL RESULTS AND DESCRIPTIVE STATISTICS

4.1. Crowd Judging

As detailed in Table I, with 7,059 units each of 8 documents, we collected more than 50,000 judgments in total, at a cost of around $\$1,700$ (CrowdFlower fees included). This is in the order of magnitude of $\$0.4$ for each document, which is broadly competitive when compared with the cost of TREC assessors or other similar crowdsourcing initiatives; for example, according to Alonso and Mizzaro [2012, Footnote 2], gathering ordinal scale judgments cost around $\$0.1$ per document. However, the comparison is more favorable when considering that in our experiments 10 redundant judgments per document are collected, and the $\mathtt{N}_k$ and $\mathtt{H}_k$ documents receive multiple judgments, whereas in Alonso and Mizzaro's work only 5 redundant judgments were collected, and there was nothing similar to our $\mathtt{N}_k$ vs. $\mathtt{H}_k$ check. We return to the cost issues in more detail in Section 6.3, where the number of repeat judgments that are required for stable ME estimates is investigated.

In total, 1481 workers participated in our experiments. Since it was impossible for a worker to perform the task twice (or more) for each topic, each worker could complete between 1 (around 30% of the workers did so) and 18 (less than 1% of the workers) tasks. As is usual in similar crowdsourcing experiments, the distribution of the number of tasks for a worker resembles a power law, although with only 18 tasks at maximum it is difficult to be precise on that. Finally, workers tend to work in sessions: when they complete a topic they either leave after that, or sometimes they complete a subsequent topic straight away—they much more rarely leave and come back later to do another.

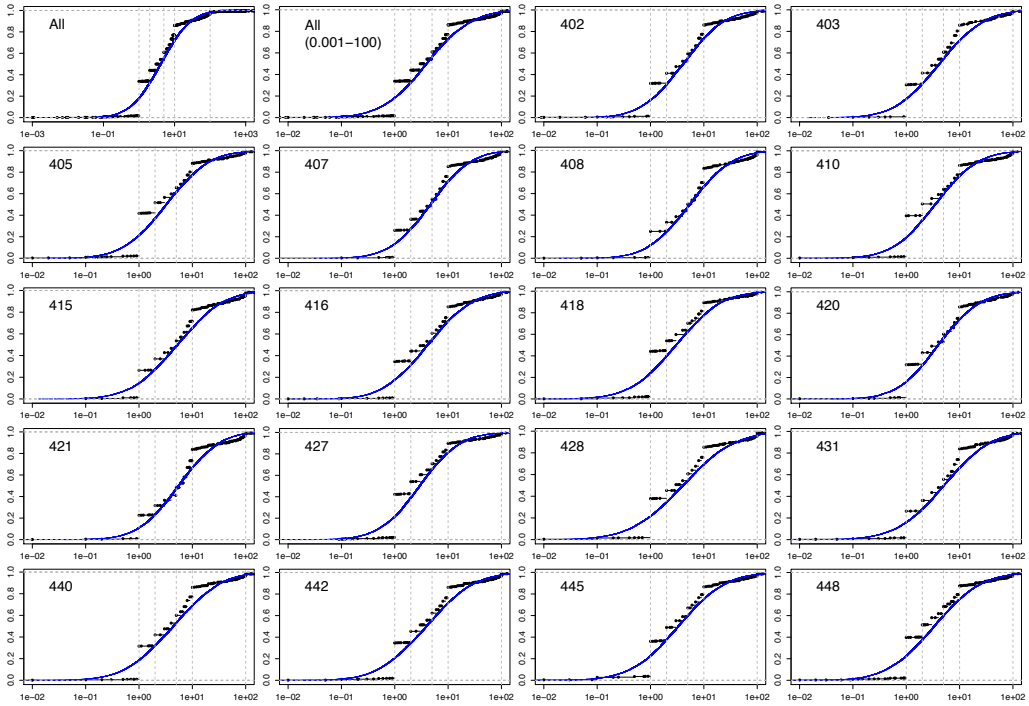

Fig. 1. Cumulative distribution functions of raw ME scores for each topic (labelled in the top left of each panel) for all scores in the range 0.01 to 100. Black dots are gathered scores and the blue line in each panel is a fitted log-normal curve. The x axis is a log scale. Vertical dashed lines are at 1, 2, 5 and 10.

4.2. Score Distribution

Magnitude estimation scores should be approximately log-normally distributed [Marks 1974; Moskowitz 1977]. Figure 1 shows the situation for our data, plotting the Cumulative Distribution Functions (CDFs) of: all the scores (top left), all the scores in the 0.001-100 range (which includes about 99.5% of the scores), and the scores for each topic. The continuous blue lines are fitted log-normal distributions with the same mean and standard deviation of the corresponding data. From the figure it is clear that the log-normal distribution is a reasonable approximation of the raw data. Differences are generally due to a *round number tendency* [Moskowitz 1977]: in general, people expressing magnitudes using a numeric scale tend to use some specific values more than others, for instance integer numbers and powers of ten. From the figure it can be seen that our workers tended to use especially 1, 2, 5, and 10 (the values where "steps" can be seen on the CDF curves — see the gray vertical dashed lines on the charts), and also quite a lot of 3, 4, 5, 6 7, 8, 9, 20, 30, 40, 50, 60, 70, 80, 90, and 100. The breakdowns on single topics show that topics do have an effect (for instance, in topics 418 and 421 the scores up to 1 included account for about 44% and 23% of the values respectively), although the general trend is very similar, as is again demonstrated by the fitted log-normal distributions. Topic means ranged from 7.9 to 12.9, and topic standard deviations ranged from 17.5 to 23.2.

Given that the scores for document relevance that we have observed are distributed similarly to other tasks where magnitude estimation has been used, we can proceed to adopt normalizing and aggregating approaches that have been used previously in the literature, as explained in the next section.

4.3. Score Normalization and Aggregation

Magnitude estimation is a highly flexible process, with observers being free to assign any positive number, including fractions, as a rating of the intensity of a presented stimulus. The key requirement is that the ratio of the numbers should reflect the ratio of the differences in perception between stimulus intensities. As the stimuli are presented in a randomized order, and observers are free to assign a number of their choice to the first presented item, it is natural that different participants may make use of different parts of the positive number space. Magnitude estimation scores are therefore normalized, to adjust for these differences. Geometric averaging is the recommended approach for the normalization of magnitude estimation scores [Gescheider 1997; McGee 2003; Moskowitz 1977], and was applied in our data analysis. Recall that for a given search topic, participants made magnitude estimation judgments for groups of eight documents (a unit). To normalize the scores, first the log of each raw score, s_i, is taken. Next, the arithmetic mean of these log scores is calculated for each *unit* of 8 documents, and for the *topic* as a whole. Each individual log score is then adjusted by the unit and topic means, and the exponent is taken of the resulting quantity, giving the final normalized score

$$s'_i = e^{\log(s_i) - \mu(\log(unit)) + \mu(\log(topic))} \tag{1}$$

An alternative equivalent formalization is:

$$s'_i = s_i \cdot \frac{\gamma}{\gamma_u}, \tag{2}$$

where γ_u is the geometric mean for each unit and γ is the geometric mean for the topic as a whole. The log transformation is theoretically motivated by the fact that magnitude estimation scores of perceived stimulus intensities have been found to be approximately log-normal, which was the case in our data as shown in Section 4.2. Intuitively, normalization by geometric averaging means that the raw magnitude estimation scores are moved along the number line, both in terms of location and spread, and the individual differences in scale are nullified. For example, as shown by both Moskowitz [1977] and McGee [2003], if two judges assigned scores of $\{1, 2, 3, 4\}$ and $\{10, 20, 30, 40\}$ to the same set of four items, respectively, then the normalized values would be identical in the two cases, as can be easily verified by applying either Equation (1) or (2). Importantly, normalization through geometric averaging has the property of preserving the ratios of the original scores, the essential feature of the magnitude estimation process.

Other normalizations have been proposed, for example using the median or the arithmetic mean instead of the geometric mean (especially when zero is an allowed score, so the geometric mean cannot be used) [Moskowitz 1977]. Other approaches include carrying out an additional calibration task that is aimed at understanding the differences in individual scales; in our scenario, this might be implemented by exploiting the $\mathtt{N}_k$ and $\mathtt{H}_k$ documents, or in a more implicit way by relying on the maximal and minimal scores expressed by each worker.

Unless otherwise noted, all analysis reported in this paper uses magnitude estimation scores normalized by geometric averaging. Geometric averaging normalization is in principle adequate since, as we have discussed, our scores are approximately log-normal and they do not include zero. Figure 2 shows the distribution of scores after normalization which are similar to Figure 1. It can be noted how the normalization has the effect of smoothing the scores (the round number tendency disappears, since each score is moved by a quantity which is different for each worker) and of making them even more log-normal.

The problem of aggregating several ME scores for the same item also needs some attention. For each topic-document pair we collected at least 10 ME scores, by 10 different workers (many more for the $\mathtt{N}_k$ and $\mathtt{H}_k$ documents). In the following we will primarily use the *median* of the normalized ME scores for a topic-document pair, although again alternatives are available at this step, such as using the arithmetic mean or the geometric mean of obtained ME scores [Moskowitz 1977]. We briefly examine the effect of different normalization and aggregating schemes on our findings in Section 6.1.

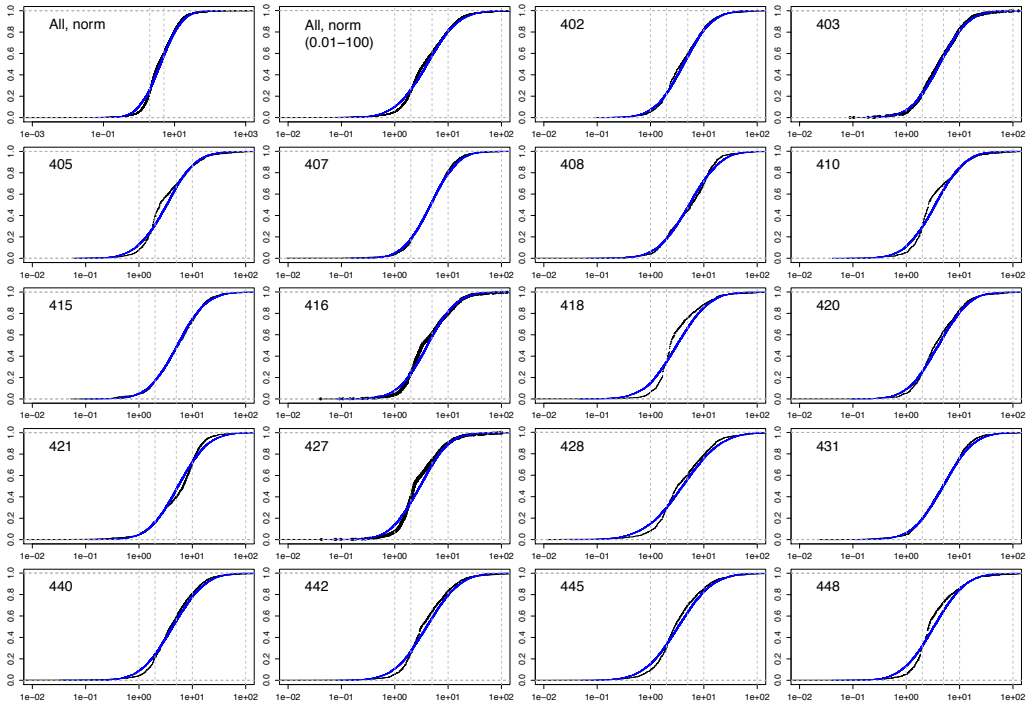

Fig. 2. CDFs of normalized ME scores using the same style as Figure 1.

5. MAGNITUDE ESTIMATION RELEVANCE JUDGMENTS

Having obtained magnitude estimation scores for 4,269 topic-document pairs, we analyze the user-perceived relevance to answer the first research question RQ1, whether the magnitude estimation technique is suitable for gathering document level relevance judgments, and whether the resulting relevance scales are reasonable with respect to our current knowledge of relevance judgments.

The distribution of the median normalized magnitude estimation scores for each document are shown in Figure 3, aggregated across all 18 topics, and split by Sormunen ordinal relevance levels (left side of figure, levels are N, M, R, H, and U, the group of documents that were not judged by Sormunen), and by TREC binary levels (right side of figure, levels are 0, 1, and U, the documents that were not judged by TREC, as not all participating runs were included in the judging pool). Boxplots in this paper show the median as a solid black line; boxes show the 25th to 75th percentile; whiskers show the range, up to 1.5 times the inter-quartile range; and outliers beyond this range are shown as individual points.

There is a clear distinction between each of the four adjacent Sormunen levels (two-tailed t-test, $p < 0.002$), with the magnitude estimation scores on average following the ordinal scale rank ordering. The differences between the two TREC levels are also significant ($p < 0.001$), with the magnitude estimation scores on average again being aligned with the binary levels. This is strong evidence for the overall validity of the magnitude estimation approach.

Figure 4 shows the magnitude estimation score distributions for each of the 18 individual topics. Although there is some variability across topics, overall the figure confirms that the magnitude estimation scores are generally aligned with ordinal categories even when considering individual topics: the medians of the median magnitude estimation scores (the solid black lines) generally follow the ordinal categories, for all categories and for all topics (there are a small number of exceptions for some of the Sormunen adjacent categories in topics 403, 405, 410, 420, 421, and 440; there are no

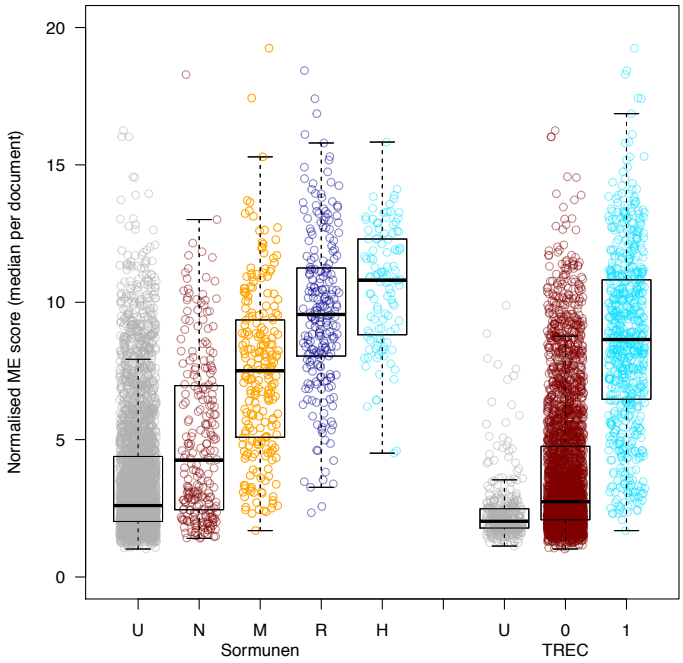

Fig. 3. ME score distribution by Sormunen and TREC levels.

exceptions for non-adjacent categories, nor for TREC categories). Since for each topic there could potentially be 3 exceptions for adjacent categories, and 6 exceptions in general, plus one exception for the two TREC categories, the 6 exceptions found are out of $3 * 18 + 18 = 72$ possible cases when considering only adjacent categories, or out of $6 * 18 + 18 = 126$ cases when considering also non-adjacent categories. Such a limited fraction of exceptions (in the 5%-8% range) is further strong evidence for the validity of our approach, even at the single topic level.

Regarding the set of documents that were not judged by Sormunen (left-most boxes in Figure 3 and sub-plots in Figure 4), based on the magnitude estimation scores it can be inferred that the bulk of this class are likely to be non-relevant; however there are also instances that occur across the central parts of the marginal to highly relevant score distributions. There were also a handful of documents unjudged by TREC that seemed to be rated highly by our judges. While the overall distributions of magnitude estimation scores are strongly consistent with the ordinal and binary categories, there are also documents in each class where the ME scores fall into the central region of a different class.

6. MAGNITUDE ESTIMATION AND CROWDSOURCING

We now turn to the second research question RQ2 and investigate whether crowdsourcing is a valid methodology to gather magnitude estimation scores. We have already touched upon several related issues: the quality checks that we put in place in our experiment and that allowed us to gather good quality data (Section 3.2.5); the approximately log-normal distribution of our magnitude estimation scores (Section 4.2); the normalization and aggregation functions (Section 4.3); and the overall agreement of the magnitude estimation scores with official relevance judgments (Section 5). We now present more details, focusing on measuring judge agreement, on analyzing disagreement with official judges, and on discussing what happens when the number of collected judgments decreases.

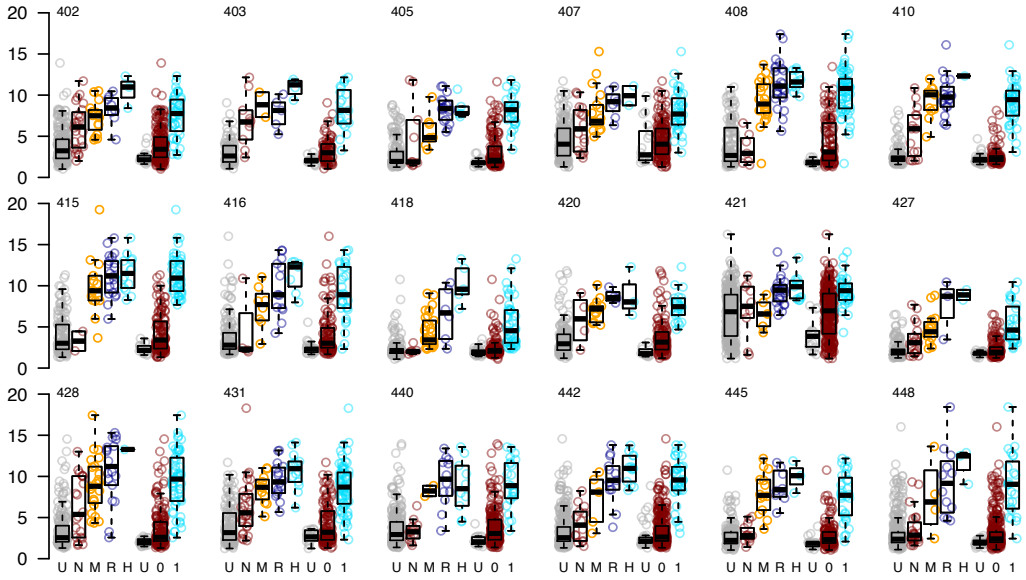

Fig. 4. ME score distribution by Sormunen (U, N, M, R, H) and TREC (U, 0, 1) levels: breakdown for individual topics.

6.1. Judge Agreement and Quality

It is well-known that relevance is subjective, even when focusing on "topical" relevance as is typically done in evaluation campaigns such as TREC, and that judges will therefore not perfectly agree. We define and discuss two kinds of agreement:

— *internal*: agreement within one group of workers judging the same topic-document pair; and
— *external*: agreement with expert judges (TREC, Sormunen).

6.1.1. Internal Agreement. In our data, there are at least ten workers judging the same topic-document pair, and the source of internal disagreement might be threefold: (i) the arbitrary, personal scale used by a judge when expressing an ME score; (ii) the subjective nature of relevance; and (iii) the often-feared low quality of work in crowdsourcing exercises. The first has already been discussed and is removed by the geometric average normalization process (Section 4.3). The second is a feature that one might not want to remove from the data: different judges can truly have different opinions on the relevance of a document, whatever scale of measurement is used. Usually IR evaluators assume that this variation is distilled into a single value (for example: mean, median or mode) that represents a population of users (although there might in fact not be a unique truth [Aroyo and Welty 2015]). Of course, the third source is in need of particular attention: low quality workers might express unreliable scores, and in general this would lead to low internal agreement.

To examine internal agreement, we can investigate the spread of the 10 normalized scores collected for each topic-document pair. Being on a ratio scale, an appropriate way of quantifying the upper limit of the spread is by the ratio between maximum and minimum $\max(s'_i)/\min(s'_i)$. Figure 5 shows the distribution of the ratio in our data. Another measure of dispersion that can be used is the geometric standard deviation (a version of standard deviation adapted for geometric mean and log-normal distributions). The chart on the right of the figure shows that the geometric standard deviation correlates quite well with the max/min ratio (Pearson's correlation is 0.999), and therefore is an almost equivalent measure.

While it is difficult to make conclusions about the absolute values of the ratio or geometric standard deviation without some reference point, it does highlight, for example, that there are 23 docu-

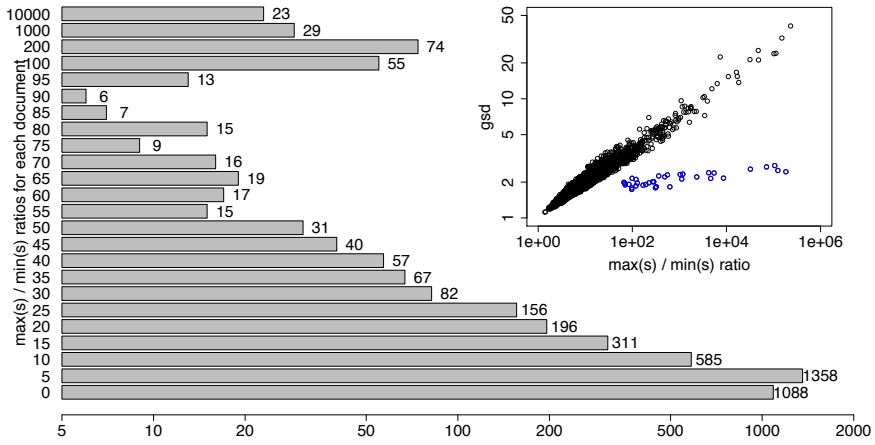

Fig. 5. Number of documents with a $\max(s_i)/\min(s_i)$ ratio in a particular range (rounded to the nearest 5), out of 4269 documents. The inner panel shows a scatterplot of the correlation between $\max(s_i)/\min(s_i)$ ratio and the geometric standard deviation (log scale; the blue dots on the bottom are the 34 out of the 36 $\mathtt{N}_k$ and $\mathtt{H}_k$ documents, which received many more judgments and therefore exhibit a higher ratio; 7 points are left out, including the two missing $\mathtt{N}_k$/ $\mathtt{H}_k$ documents).

ments that seem to have an unusually high ratio of 10000 or more. Looking at each of these cases, there are two distinct causes. Firstly, 15 of the 23 looked like genuine attempts to assign ME scores, but the range of scores for each worker involved was very wide. One reason given in the comments by workers was that there was no zero available, so for documents that were truly off topic, they chose an arbitrary, very very small, number, making a wide scale. These wide scales were not markedly compressed by the normalization stage. In some cases the numbers assigned seemed arbitrary, but suitably large or small in correlation with the expert judgments. The second cause, in 8 of the 23 cases, seemed to be workers not following instructions or reading documents carefully. In these cases the comments provided by the workers bore no relationship to their numeric scores, and the numbers themselves seemed arbitrary. We will discuss the seemingly arbitrary nature of some scores later on (Section 9.2).

Another way of examining internal (intra-assessor) agreement is through measures of reliability. Krippendorff's α [Hayes and Krippendorff 2007] is a widely-used reliability measure that is defined for different types of measurement scales. α is a chance-corrected measure of reliability, taking into account both observed and expected agreement. For ratio-level data, the α difference function is defined as square of the ratio of the difference between a pair of values assigned by two judges to the sum of the values assigned by the same pair of judges [Krippendorff 2011].

Across the full set of 4,269 topic-document pairs, $\alpha = 0.323$ (recall that each item received at least 10 independent judgments; for the few items where a larger number of judgments were collected, the first 10 responses are used). Based on a bootstrap hypothesis test, this level of agreement is significantly different from zero agreement ($p < 0.01$). We note that the numeric value of α is difficult to interpret in this context, as to the best of our knowledge this is the first investigation of document-level relevance judgments using magnitude estimation, so no direct comparison is available.

Overall, the data show a not perfect but reasonably high internal agreement.

6.1.2. External Agreement and Quality of Workers. We now turn to analyze the external agreement, namely the agreement with the "official" TREC and Sormunen judges. In passing we also investigate the level of agreement when using different scales for judging relevance. We have al-

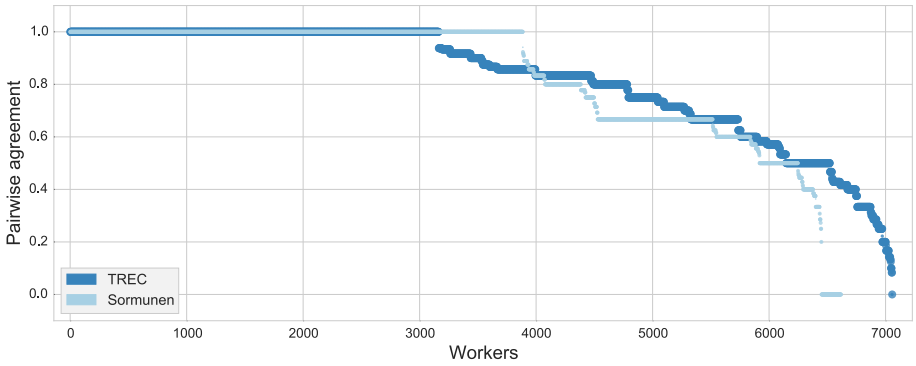

Fig. 6. Agreement over the ranking of all pairs of documents for individual units using worker-assigned ME scores versus TREC or Sormunen categories. The number of pairs for each unit may vary, since some of the documents might be unjudged on the ordinal scales.

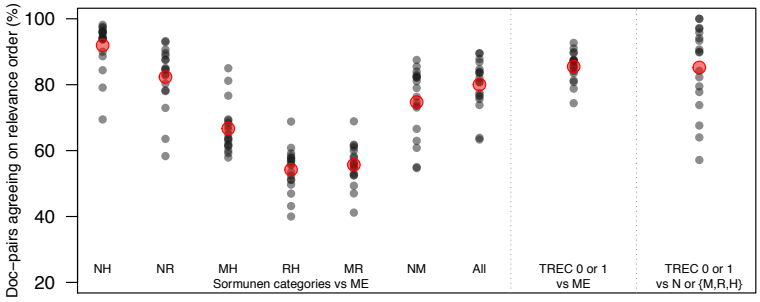

Fig. 7. Raw agreement on the ordering of relevance of all document pairs (one small dot per topic) between judges as indicated on the x-axis (ME: magnitude estimation). The large (red) dot is the mean over all topics in each case. Unjudged documents are excluded.

ready discussed the overall consistency with the official judges in Section 5. Here the results of a more specific analysis are shown: comparing the pairwise orderings between binary, ordinal and magnitude estimation relevance judgments. For example, if two documents are rated N and M on the ordinal scale, and the corresponding magnitude estimation score of the first is lower than or equal to the score for the second, this would be counted as an agreement. If the magnitude estimation score was higher for the second document, it would indicate disagreement. Intuitively, we can estimate that a worker has a high quality if the number of pairs of documents are ordered as in TREC or Sormunen. This measure of the quality of our crowdsourcing workers is reasonable even if the scales of the judgments differ, as long as they impose a ranking on the documents.

Figure 6 shows, for each unit, the pairwise agreement with both TREC and Sormunen judgments. Workers are generally good using this metric, with around 50% of the workers having perfect agreement; around 70% an agreement higher than 0.8; and around 86% of the workers have an agreement higher than 0.6. Note that as there are many documents unjudged on the Sormunen scale, thus 2376 units have only one pair of documents on which to compute agreement, contributing to the appearance of workers agreeing more with Sormunen than TREC using this metric.

Figure 7 shows the agreement data stratified by each of the possible six pairs of ordinal relevance levels of the Sormunen judgments (first six columns) and ungrouped (seventh column). The TREC vs. ME comparison, aggregated over all topics, is shown in the eighth column, and for comparison, the final column shows the pairwise agreement between TREC and Sormunen, assuming that N equates to TREC category “0”, and the other three Sormunen categories equate to TREC category

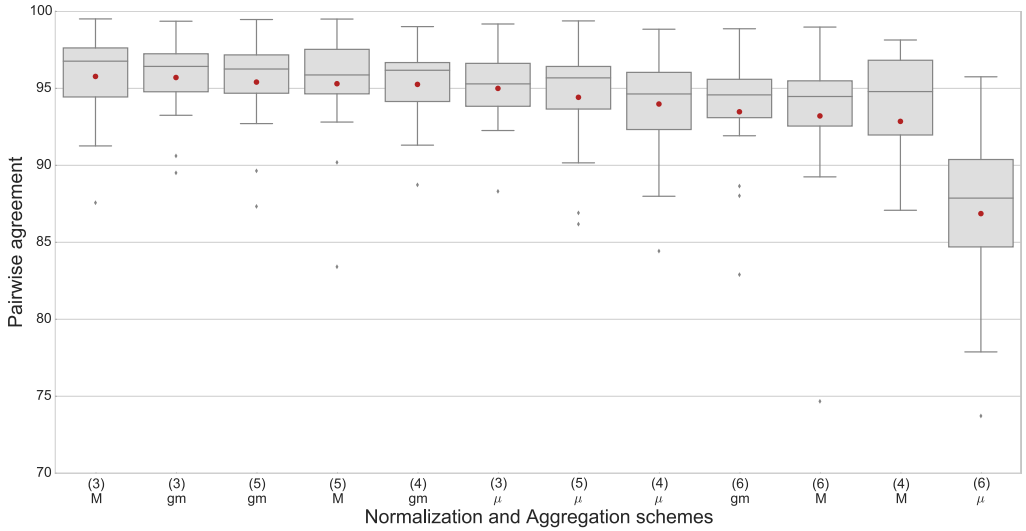

Fig. 8. Effect of the four normalization functions (Equations (3), (4), (5), and (6) in the text) and aggregation functions (M = median, gm = geometric mean, and μ = arithmetic mean) on pairwise agreement.

"1". Red large circles indicate the mean score over the 18 topics for each group. It can be seen that the rates of agreement are highly consistent when comparing any of the three relevance scales. In particular, ME scores for H documents are greater than those for N documents 92% of the time (mean, first column), which is higher than the average agreement between TREC and Sormunen (85%—mean, last column). Similarly, the proportion of document pairs that have different TREC categories are ranked in the same order by ME scores 86% of the time (mean, second last column). ME and Sormunen ranking for documents in the NR category also agree 82% of the time (mean, second column), and over all pairs Sormunen and ME agree 80% of the time (mean, seventh column). This overall average is reduced by the lower agreements between pairs of documents that are deemed relevant and one or other are in the M or R categories (columns three to six). This is perhaps unsurprising, as distinguishing "marginal relevance" (M) from "irrelevant" (N) or "relevant" (R) can be difficult; and likewise distinguishing R from highly relevant (H) can also be challenging.

Overall, the agreement between the ME approach and the existing categorical judgments seem reasonable, further supporting the validity of the magnitude estimation approach for gathering relevance judgments.

We now revisit the normalization and aggregation issues discussed in Section 4.3. We compare the following four normalization schemas

$$s'_i = e^{\log(s_i) - \mu(\log(unit)) + \mu(\log(topic))} \tag{3}$$

$$s'_i = e^{\log(s_i) - M(\log(unit)) + M(\log(topic))} \tag{4}$$

$$s'_i = e^{\log(s_i) - \mu(\log(\max(unit)), \log(\min(unit))) + \mu(\log(topic))} \tag{5}$$

$$s'_i = e^{\log(s_i) - \mu(\log(\mathtt{H}_k), \log(\mathtt{N}_k)) + \mu(\log(topic))} \tag{6}$$

(where the first one corresponds to that already discussed in Equations (1) and (2)), and the three aggregation functions mentioned at the end of Section 4.3, namely median, geometric mean, and arithmetic mean. Figure 8 shows the effect of the different normalization and aggregation functions on the quality of the workers measured by pairwise agreement. As we have discussed in Sections 4.2 and 4.3, we decided to use the somehow standard geometric averaging (Equation (3)) plus median normalized score. The chart in figure clearly shows that the selected two functions are among those

avoiding a loss of quality. Both when considering the medians in the boxplots and when considering the means (the red dots in figure), on our data the highest pairwise agreement is obtained with:

— the normalization scheme in Equation (3), and
— the median or the geometric mean aggregators, rather than the arithmetic mean.

6.2. Failure Analysis

Despite the similar overall levels of agreement between the magnitude estimation method and ordinal relevance, Figures 3 and 4 show that some individual documents appear to be "misjudged". We therefore conducted a failure analysis, manually examining a subset of documents for which the Sormunen relevance level and the median magnitude estimation scores were substantially different (for example, where a particular document was assigned an ordinal Sormunen relevance level of `N`, but the median magnitude estimation score for the document was closer to the magnitude estimation scores assigned to `H` documents for the same topic, and substantially higher than the magnitude estimation scores assigned to other `N` documents for the same topic). Based on the manual examination of 34 documents where there appeared to be a significant flip between the ordinal and magnitude estimation scores, we found two broad classes of disagreements: those where one group of assessors appeared to be clearly wrong; and a class where the topic statement itself is so unclear as to be open to interpretation.

Of 34 documents that were examined, we found 14 cases (41.2%) where the Sormunen ordinal judgments appeared clearly wrong, and 9 (26.5%) cases where the crowd-based magnitude estimation assessments appeared clearly wrong. For this class of clear disagreements, where some assessors appear to be clearly wrong in the assignment of relevance (whether ordinal or magnitude estimation), the cause mostly appears to be that the assessors have missed or ignored a specific restriction included as part of the TREC topic. For example, the narrative of topic 410, *"Schengen agreement"*, includes the statement that: *"Relevant documents will contain any information about the actions of signatories of the Schengen agreement such as: measures to eliminate border controls..."*. Document `FT932-17156` makes clear reference to nine signatories of Schengen, and the process of removing passport checks. As such, it seems implausible that the document should be classed as `N`, or completely non-relevant. The original TREC binary judgment supports this view, having assigned a rating of 1 (indicating that the document is at least marginally relevant).

For the remaining 11 (32.4%) cases, it was not possible to determine that one assessment was clearly correct and the other wrong. Here, the original TREC topic statement itself was ambiguous, preventing a clear conclusion to be drawn based on the limited information that the topic statement provided. For example, a number of topics list several concepts in the narrative about what is or is not deemed relevant. However, they introduce ambiguity about whether the document must meet all of the listed criteria, or whether a subset is sufficient. For example, topic 407, *"poaching, wildlife preserves"*, states that *"A relevant document must discuss poaching in wildlife preserves, not in the wild itself. Also deemed relevant is evidence of preventive measures being taken by local authorities."* This raises the ambiguity of whether preventative measures by authorities against poaching, but not specifically in wildlife preserves, should be considered as being at least somewhat relevant, or completely non-relevant. We note that further ambiguity is introduced due to the temporal mismatch between the time when the documents and topics were written (1990s), and when the magnitude estimation judgments are being made (2010s). This is particularly the case for topics that include terms such as "current".

The above failure analysis must also be interpreted in the context that it is known that assessors make mistakes when judging, perhaps due to fatigue or other lapses in attention, leading to self-inconsistencies [Carterette and Soboroff 2010; Scholer et al. 2011]; or they may display systematic errors due to a misunderstanding of the relevance criteria, or relevance drift [Webber and Pickens 2013]. Clearly, assessor errors will lower overall agreement rates when comparing assessments. Determining whether magnitude estimation relevance assessments lead to higher or lower error rates compared to using ordinal or binary scales is left for future work.

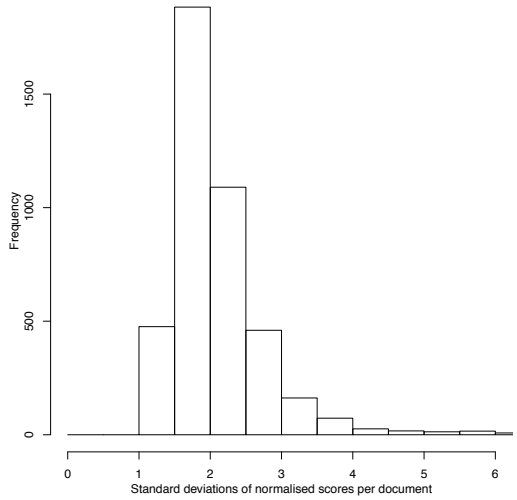

Fig. 9. Histogram of standard deviations of the normalized scores for each topic-document in our data. There were 53 values greater than 6 which are not shown.

Overall, the examination of a set of clear disagreements demonstrates that there are cases where both groups of assessors (ordinal or magnitude estimation) are at odds with certain details of the TREC topic statements, and that these appear to occur at broadly similar rates. Moreover, the topic statements themselves are sometimes a cause of ambiguity, placing a practical upper-limit on the agreement that can be achieved. We conclude therefore that the magnitude estimation relevance judgments are sound and sensible, even when collected by means of crowdsourcing, having similar agreement rates with the ordinal Sormunen judgments as the Sormunen judgments have with TREC assessments.

6.3. How Many Workers are Required?

In Section 6.1.1 we observed that there is a spread of scores assigned by different workers to the same document for a topic, and that this is reasonable given that relevance is subjective quantity. For system evaluation it is typical to distill this variation into a single aggregate measure of relevance such as the mean or the median. This value is assumed to be an estimate of the true population relevance score for the document. If we have an estimate of the population variance, we can use standard sample size calculations to determine how many ME scores we should collect per document to be confident that we are estimating the true population mean score accurately. Specifically, to be 95% confident that our sample mean is within E of the true mean we need

$$n = 1.96^2\sigma^2/E^2$$

samples, where σ is the population standard deviation.

While we do not know the population standard deviation for ME scores of a particular document, we can look at the distribution of sample standard deviations that we have in our data as a guide. Figure 9 shows a histogram of the standard deviations of our normalized scores for each document. Using these as a guide, it would seem that assuming σ around 2 is reasonable. Table II shows the number of workers required per document for various values of E and σ. Assuming that we can tolerate ± 2 in our estimate of the true mean relevance, then we need about 7 workers, assuming $\sigma = 2$.

While this argument is far from being conclusive, it confirms that the ten workers that we used in our experiment, and the associated cost, can be reduced. Therefore, from the about $\$0.4$ for each

Table II. The number of workers required to judge each document to be 95% confident that mean normalized scores are within E of the true population value.

	σ					
E	1	2	3	4	5	6
0.5	27	107	239	425	664	956
1.0	7	27	60	107	166	239
1.5	3	12	27	48	74	107
2.0	2	7	15	27	42	60
2.5	2	5	10	17	27	39
3.0	1	3	7	12	19	27
3.5	1	3	5	9	14	20
4.0	1	2	4	7	11	15
4.5	1	2	3	6	9	12
5.0	1	2	3	5	7	10

document figure mentioned in Section 4.1 we can estimate that only about $0.3 for each document are really needed, if not less since more efficient methods could be devised.

6.4. Summary

Using ME with crowdsourcing could potentially raise several concerns, including: subjectiveness of the relevance scale; difficulty in normalizing disparate scales from different workers; lack of proper training of workers; potential internal and/or external disagreement between workers; and high overall cost as redundant workers needed to ensure quality. On the basis of the results presented so far we can conclude that, at least with our setting and our quality checks in place (see items (a)–(d) in Section 3.2.5), we can be reassured that magnitude estimation relevance scores gathered by means of crowdsourcing are reliable, thus answering RQ2. Also, the cost to collect the ME judgments in our experiments is comparable to that of collecting ordinal scale judgments, and there are hints that this cost can be lowered by collecting fewer judgments.

7. MAGNITUDE ESTIMATION FOR SYSTEM-LEVEL EVALUATION

The third research question RQ3 concerns the direct application of magnitude estimation relevance judgments for the evaluation of IR systems, by considering their use for calibrating gain levels in two widely-used gain-based IR evaluation metrics, nDCG and ERR.

7.1. Gain in the nDCG and ERR Metrics

Magnitude estimation provides scores that reflect ratios of human perceptions of the intensity of relevance of different documents in relation to a topic. This is directly related to the notion of gain in effectiveness metrics such as normalized discounted cumulative gain (nDCG). For example, Järvelin and Kekäläinen [2002] describe "cumulative relevance gain" that a user receives by examining a search results list, and discuss setting "relevance weights at different relevance levels". That is, weights are applied to each level of an ordinal relevance scale, and can be chosen to reflect different assumptions about searcher relevance behavior. However, the "standard" approach that has been adopted when using nDCG is to simply assign ascending integer values to the ordinal levels, starting with 0 for the lowest (non-relevant) level; for a 3-level ordinal scale, the default gains would be 0, 1 and 2, as for example implemented in `trec_eval`.[1]

In addition to modeling different levels of gain, or relevance, the discounted cumulative gain metric also includes a discounting function, so that documents that are retrieved further down in a ranked search results list contribute smaller amounts of gain, reflective of factors such as the effort or time that a user must invest while working their way through the list [Järvelin and Kekäläinen

[1] http://trec.nist.gov/trec_eval

2002]. Using a logarithmic discount, discounted cumulative gain at cutoff N is calculated as

$$\text{DCG@N} = G_1 + \sum_{i=2}^{N} \frac{G_i}{\log_2(i)},$$

where G_i is the gain value for the document at position i in the ranked list. To enable fair comparisons across topics with different numbers of relevant documents, the DCG@N score is divided by an *ideal* gain vector, a ranking of documents in decreasing relevance order, to obtain normalized discounted cumulative gain, nDCG@N.

Expected reciprocal rank is a metric based on a cascade model of searcher behavior, where the probability of continuing on to the next position in the ranked results list is influenced by the relevance of previous items. The metric calculates the expected reciprocal rank at which the searcher will stop [Chapelle et al. 2009], and is defined as

$$\text{ERR@N} = \sum_{i=1}^{N} \frac{R(G_i)}{i} \prod_{j=1}^{i-1} (1 - R(G_i)),$$

where $R(G) = (2^{G_i} - 1)/2^{\max(G_i)}$. In the analysis that follows, we calculate both nDCG and ERR up to a depth of $N = 10$.

For both nDCG and ERR, gain is based on the relevance of a document, which has previously been measured on an ordinal scale, typically with 3 [Kanoulas and Aslam 2009] or 4 [Järvelin and Kekäläinen 2002] levels, depending on the test collection being used. Since magnitude estimation relevance judgments are continuous rather than ordinal in nature, and reflect the perceived level of relevance for individual topic-document combinations, it is possible to assign more fine-grained gain values, potentially reflecting different gains for individual documents, or even for individual searchers.

7.2. Comparative System Rankings

Given that the agreement between our magnitude scores, TREC judgments and Sormunen's judgments is not perfect, it is reasonable to expect that relative system effectiveness orderings computed with each of these as a basis may differ. We first examine the correlation between system orderings using magnitude estimation scores and TREC relevance as gain values, as for both these judgment sets we have nearly complete coverage of the top 10 documents of all runs submitted to TREC-8, for our 18 judged topics.

Rather than simply taking the median of all ME scores assigned to a document, we take the median of scores taken from the three units that have document orderings that have the highest agreement with the TREC orderings, or Sormunen orderings, respectively. In effect, the scores derived from units with high internal agreement as computed in Section 6.1.1, with ties broken in favor of units with the least unjudged documents. By aggregating scores over units that have a high agreement with the original TREC orderings, any differences in system rankings we see should be more attributable to the scale of the gain values, rather than perturbations in document orderings given by different gain values. Note that this is a slightly less confounded methodology than was used in our previous paper [Turpin et al. 2015], which simply took the median of all units, even if the unit wildly disagreed with the ordering of documents given by the categorical judgments. This revised methodology accounts for the differences between the figures and tables in this section and in the previous paper.

Figure 10 shows the concordance between system rankings using TREC categories and ME scores as gain values. It can be seen that there are definite changes in system rankings when using the different relevance scales, with Kendall's τ correlations of 0.677 and 0.222 for nDCG@10 and ERR@10, respectively. As the aim of system evaluation is to identify top-performing systems, we also consider changes in the *top set*, defined as the group of systems that are statistically indistinguishable from the system with the highest mean effectiveness, using a paired Wilcoxon signed-rank test with

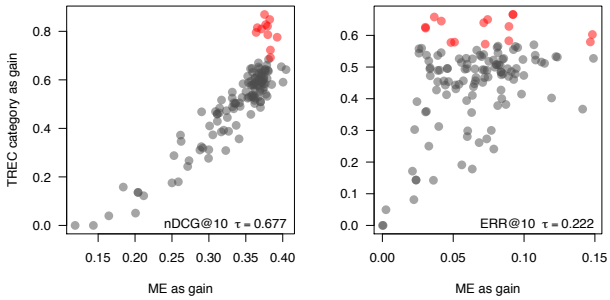

Fig. 10. System scores using TREC categories (y-axis) and magnitudes (x-axis) as gains in the nDCG@10 and ERR@10 metrics. There is one dot for each of the 129 systems participating in the TREC-8 ad hoc track, with red dots showing the top-set (see text). Magnitudes used are the median of the three units per document that most agree with TREC document rankings. Kendall's τ is shown.

$p < 0.05$. These systems are shown as red dots. The overlap of the top set, based on the TREC and magnitude estimation relevance judgments is 82% for nDCG@10 and 20% for ERR@10, confirming that the perturbation in system ordering has an impact on which systems are identified as being the best performers.

Since the Sormunen judgments only cover around 19% of the documents that occur in the top 10 of all TREC-8 ad hoc runs, evaluating the original runs using these relevance judgments is problematic, due to the large number of unjudged documents. However, the judgment coverage of individual ranked lists (that is, for particular topics within a run) varies substantially. Therefore, to enable a comparison between the ordinal judgment and magnitude estimation judgments on system orderings, we simulate runs that only include ranked lists for topics that are fully judged by Sormunen.

For each topic, we identify all ranked lists within the full set of runs that have Sormunen judgments for all top 10 documents for that topic; we call each of these a *complete sub-run*. There are 12 topics where there are at least two complete sub-runs out of all of the TREC-8 ad hoc runs for that topic. To form a simulated retrieval system, we then randomly choose one complete sub-run for each of the 12 topics to give a "full run" over all 12 topics. Using this method, we construct 100 simulated system runs that are random merges of complete sub-runs from real runs. While these simulated systems are not actual TREC submissions, they are plausible in that each individual topic ranking is from a real TREC run.

Figure 11 shows that there is also a large discordance between the system rankings using the Sormunen categories and magnitudes as gain values. However, note that the scale is different in this figure than in Figure 10, as the simulated systems all have relevant documents in the top 10, and so are high scoring. Correlation as measured by Kendall's τ is even lower here, at 0.147 and 0.06 for nDCG@10 and ERR@10, respectively. The overlaps in top sets are 32% for nDCG@10 and 68% for ERR@10. The relatively high performance of systems on this simulated collection is to be expected, as only complete sub-runs were selected, and around half of the Sormunen judgments are for documents that were originally rated as relevant by TREC, and so the runs in the simulated systems have a high number of relevant documents. It is interesting that there is still a sizable change in the top set using either metric with the different judgments, however.

7.3. Judgment Variability and System Rankings

The previous section showed that using magnitudes as gain led to changes in the ranking of systems using different metrics, even when only aggregating over magnitude judgments that agreed with individual pairwise document rankings of the original categorical scale. There is, however, substantial variation in the magnitudes assigned to documents by our judges. Perhaps using magnitudes from different judges would lead to different system rankings. As we have at least 10 judgments per topic-document pair, we can re-sample individual scores many times, recompute system orderings, and

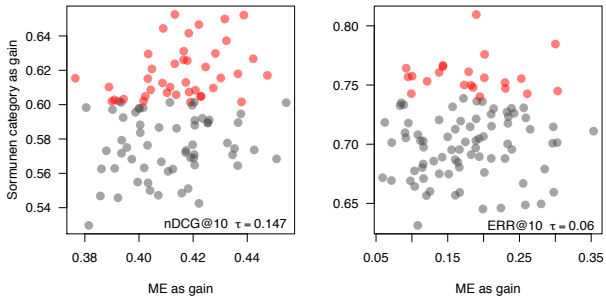

Fig. 11. As for Figure 10 but using Sormunen categories (y-axis) and simulated runs as described in the text.

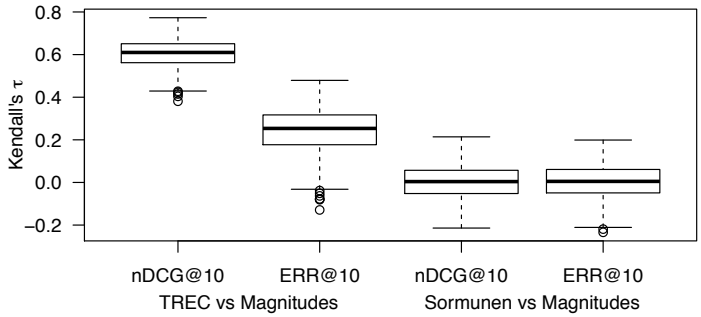

Fig. 12. Kendall's τ between system rankings obtained using gains from the judgments and metrics indicated on the x-axis, where magnitudes are randomly sampled for each document 1000 times. Compare with τ in Figures 10 and 11.

get a distribution of τ values. The results are shown in Figure 12: the correlations between system orderings seen in the previous section sit around the middle of these distributions. The spread of the distributions reflect the wide range of individual variation – differences in perceptions of relevance – that in turn leads to a wide range of τ values.

An alternate explanation for the low τ values is that gains set as magnitudes are generally much higher than the gains set by Sormunen's categories. From Figure 3 it is apparent that magnitudes are generally in the range of 2 to 15 (although some are as large as 100 or 1000), whereas using categories as gains the values are 0, 1, 2 or 3. However, because both nDCG and ERR are normalized, this scale effect is nullified to a degree. For example, multiplying gains by a constant has no effect on nDCG, and even altering the category scores using a small exponential or additive constant has little effect. Table III shows the effect of using different methods for converting the categorical values (C_i) from Sormunen into gain values (G_i) for computing nDCG and ERR using our 12 topics on the simulated systems. Kendall's τ is computed between the system rankings given using the categorical values as gains ($C_i = G_i$), and the other methods, hence the top row has perfect correlation. The final two rows hint that using gain values from a broad scale, while preserving document orderings, may alter system rankings, particularly with nDCG@10, compared with simply using the category values as gains. The final row perhaps best represents our magnitude scores, and so goes part way to explaining the low τ values in the previous section. It does not, however, explain all of the changes in system rankings: it still seems that our ME judgments are different in more than scale.

One way to examine this further is to split our data into *narrow* units, those whose $\mathtt{H}_k/\mathtt{N}_k$ ratios are less than 5 (the median value), and *wide* units, where the ratios $\mathtt{H}_k/\mathtt{N}_k$ are 5 or more. Figure 13 shows a histogram of the number of units with each ratio. This creates a set of judgments with a

Table III. Correlation (Kendall's τ) and overlap of top-set between system orderings when using Sormunen categories as gain ($G_i = C_i$), and other mappings between categories and gains on the simulated systems, when using the nDCG@10 and ERR@10 metrics.

	Kendall's τ		Top-set overlap	
Mapping	nDCG@10	ERR@10	nDCG@10	ERR@10
$G_i = C_i$	1.000	1.000	100%	100%
$G_i = 100 \times C_i$	1.000	0.698	100%	100%
$G_i = 1 + C_i$	0.989	0.969	98%	95%
$G_i = 10 + C_i$	0.961	0.939	95%	91%
$G_i = 100 + C_i$	0.957	0.939	95%	100%
$G_i = 10^{1+C_i}$	0.720	0.698	66%	100%
$G_i = 2^{1+C_i}$	0.870	0.702	42%	73%

Table IV. Correlation (Kendall's τ) and overlap of the top set between system orderings when median document magnitude scores are based on narrow, wide or all units.

Source of gains	Narrow	Wide	All
TREC, τ			
nDCG@10	0.679	0.744	0.677
ERR@10	0.361	0.247	0.222
TREC, top set			
nDCG@10	63.64%	72.73%	81.82%
ERR@10	20.00%	60.00%	20.00%
Sormunen-Simulated, τ			
nDCG@10	0.036	0.382	0.147
ERR@10	0.017	0.323	0.060
Sormunen-Simulated, top set			
nDCG@10	9.76%	68.29%	31.71%
ERR@10	68.18%	100.00%	68.18%

scale similar to Sormunen (the narrow units), and a set of judgments that are over a larger range (wide units).

Table IV shows the correlation (Kendall's τ) between system orderings, and overlap of runs in the top set, when using TREC and Sormunen judgments, and magnitude estimation scores when the median is obtained from units that used a narrow scale, a wide scale, or from all units (note that the τ values in the All column match those in Figures 10 and 11). In each case the score for a topic-document pair is taken as the median of the 3 units that most agree with TREC or Sormunen and that are either narrow, wide or any as appropriate. When there was no narrow or wide judgment for a topic-document pair (all units containing that topic-document pair had $\mathtt{H}_k/\mathtt{N}_k$ above or below the median), the median of the three best units was used. This occurred in 12% of TREC topic-document pairs, and 7% of Sormunen topic-document pairs.

Generally the correlations between the system rankings based on categorical and ME gain values are higher when the wide judgments are used as opposed to the narrow judgments. This contradicts our expectations from Table III, which indicated that a larger scale of gains led to a decrease in correlations. This confirms that a simple argument based on the scale of judgments is not sufficient to explain the different system orderings, although the number of topics used here is small (12). We examine this further in the next section.

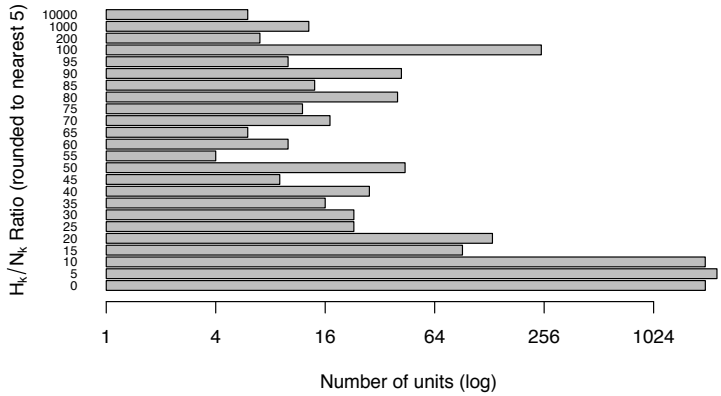

Fig. 13. Number of units with a particular $\mathtt{H}_k/\mathtt{N}_k$ ratio over all topics. There were a total of 7,059 units. 37 units with frequency less than 4 are not shown.

7.4. Summary

In answering RQ3, we have shown that there are clear differences in system rankings using nDCG and ERR metrics on this data when using ME scores as gain values compared with using ordinal categories as gain values. Depending on the judgments and the metric used, correlations between system orderings using ME and the categorical judgments, as measured with Kendall's τ, varied between about 0.8 and -0.2. At first this might seem like a simple scaling problem, where ME judgments are over a wider range than categorical judgments. However, analysis shows that simple factor cannot be the whole answer; a much more likely explanation is that there is substantial individual variance in not only how people perceive the relevance of an individual document, but also in how they perceive the ratio of relevance of one document to another.

8. INVESTIGATING GAIN USING MAGNITUDE ESTIMATION

The previous section used perturbations in system orderings to examine the difference between using magnitude estimate relevance scores and the usually employed ordinal relevance levels as gains. This section directly considers gain profiles, answering our fourth research question RQ4 by considering what additional insights magnitude estimation can provide into user perceptions of relevance.

The gain weights of the nDCG metric allow for the modeling of different user relevance preferences. However, in practice gains are usually set to one of two profiles: a *linear* setting (as defined in Section 7); or an *exponential* setting, where the ordinal relevance level forms a power of 2 [Burges et al. 2005], placing more emphasis on highly relevant documents. Figure 14 shows the distribution of magnitude estimation scores, normalized by the $\mathtt{N}_k$ score in each unit (intuitively, a measure of relevance standardized for each user). Black lines show the median magnitudes for each group. Superimposed are the two default profiles, linear (white circles), and exponential (white triangles). It can be seen that, compared to actual user perceptions of the relevance space, the linear gain profile is fairly close to the gain that should be allocated at each level to satisfy the (mythical) median user. The exponential profile is further from the median, overestimating the gain for the R and H levels, but catering for some judgments.

It is readily apparent, however, that the large spread of the distributions of magnitudes cannot be captured in a single gain formulation. There is simply too much individual variation to warrant the simplification of relevance perceptions to a single profile.

Kanoulas and Aslam [2009] investigated how gain values could be set in nDCG so as to maximize the stability of system rankings, based on the generalizability and dependability coefficients from Generalizability Theory. The settings were estimated empirically, based on system runs from

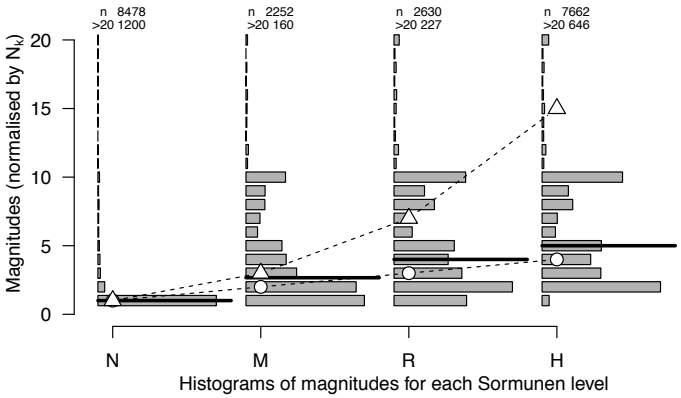

Fig. 14. Distribution of magnitudes (normalized by N_k) for each Sormunen level. Circles are linear gain values, $\{1, 2, 3, 4\}$, while triangles represent exponential gain values, $\{2^1, 2^2, 2^3, 2^4\}$. Black lines are medians; the text "n" gives the total number of magnitudes in the level, and "> 20" shows the count of scores above 20.

the TREC 9 and 10 Web Tracks and the TREC 12 Robust Track. We compare this approach with ME scores of relevance perceptions gathered from users. Figure 15 shows the ratio of magnitudes for document pairs in the Sormunen categories highly relevant and partially relevant. The ratios are computed within each unit collected (8 documents for one topic), and if the unit contains more than one document in some category, all combinations of the documents are included in the graph. To allow comparison with gain values recommended by Kanoulas and Aslam [2009], we either ignore marginally relevant documents (Sormunen category M), leftmost box; fold them in as partially relevant, middle box; or call them partially relevant and promote category R to highly relevant. When a unit did not contain a marginally relevant document, it is omitted from the middle and right boxes, hence there are less grey dots for those plots.

Kanoulas and Aslam [2009] recommend a ratio of 1.1 for the gain values of relevant to highly relevant documents to maximize the effectiveness of a system comparison experiment. Using magnitudes directly as gain values gives a median of 1.00, 1.00, and 1.14 in the three boxes in Figure 15, which happens to be close to their recommendation. However, again, we see a wide variation from this median, and mean values of the ratios are 48, 22 and 24 respectively, further suggesting that there may not be a single gain setting that is suitably representative across different user perceptions and tasks.

In the previous section, and in this section, we have seen not only that relevance values for a single document vary among users, but the ratio of relevance between two documents also varies. Thus, in answer to RQ4, the selection of a single gain formulation as either a linear or exponential mapping of categorical judgments is guaranteed to misrepresent at least half of our users in this data set. Similarly, choosing a single gain formulation to maximize the statistical efficiency of system comparisons will also misrepresent a large part of the user population.

9. CONCLUSIONS AND FUTURE WORK

Relevance is a fundamental concept in information retrieval, underpinning test collection-based system effectiveness evaluation. This is the first study to have collected large scale, real user data about relevance perceptions using magnitude estimation, obtaining through crowdsourcing over 50,000 ratings across 18 TREC topics. We close the paper by briefly summarizing our findings and sketching future work, as well as highlighting some limitations of this study.

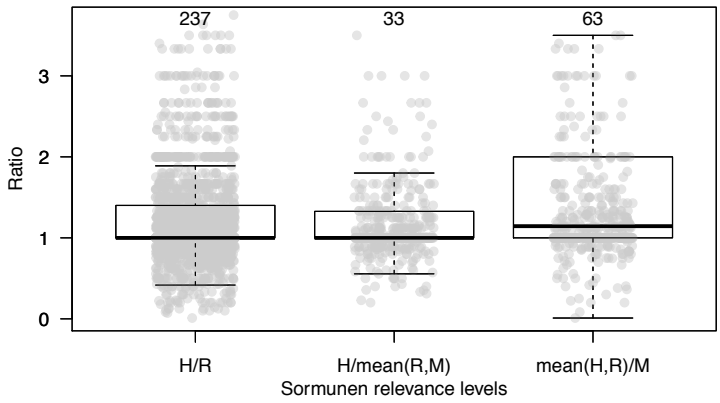

Fig. 15. Ratio of the magnitude scores assigned to the Sormunen categories highly relevant (H) to relevant (R) and marginally relevant (M) over all pairs of documents. Numbers indicate the count of document pairs that are greater than 3.5 for each box. Each grey dot is a ratio of one document pair.

9.1. Summary

The first research question that was considered asked whether magnitude estimation is a suitable technique for judging relevance at the document level. A comparison of the overall distribution of magnitude estimation judgments shows a close alignment with ordinal graded relevance categories, both at the individual topic level, and across the set of 18 topics as a whole. A high level of consistency was also established between the distribution of magnitude estimation judgments and classical binary relevance assessments. A failure analysis demonstrated that for those cases where the magnitude estimation and ordinal judgments disagree, this is typically not due to problems with magnitude estimation per se, but rather arises due to different judging context, or from ambiguity in the original topic statements.

Magnitude estimation is a technique that may be unfamiliar to many people. The second research question therefore focused on the suitability of using crowdsourcing to obtain magnitude estimation judgments at scale. Aiming to avoid spam workers, our crowd task included four quality checks: a practice line length estimation task; a multiple choice topic understanding question; a check that magnitude estimates that were entered for two documents with known ordinal relevance were rank consistent; and, a minimum time of 20 seconds spent per document. The magnitude estimation judgments that were obtained after this process showed a high level of external agreement with existing TREC binary and Sormunen ordinal judgments: rank ordered magnitude estimation scores agree with binary judgments 86% of the time, and with ordinal judgments 80% of the time (the agreement with ordinal judgments is higher for ordinal levels that are further apart, and lower for ordinal levels that are closer together). The ranked agreement between binary and ordinal judgments was 85%, giving confidence that the using magnitude estimation is robust as a scaling technique for document-level relevance.

Considering the impact of magnitude estimation relevance judgments on IR system evaluation, the third research question, our analysis showed that results can vary substantially when different relevance scales are used, with a τ of 0.677 between system orderings when effectiveness is measured using nDCG@10 and comparing magnitude estimation scores with binary relevance across TREC runs. When comparing magnitude estimation with ordinal relevance on simulated runs, the correlation is even lower at 0.147. Moreover, when individual magnitude estimation scores are considered, rather than the mean, the correlations show a wide range of individual variation in terms of the perception of relevance. A key factor appears to be that different searchers have different perceptions of the extent of relevance variation, using wide or thin scales. Overall, the analysis suggests

that it is important to incorporate different judging scales, rather that assuming that a single scale can closely reflect a larger population of users.

We also employed magnitude estimation to directly investigate gain profiles, the fourth research question. Typically, gains in nDCG are set as linear or exponential profiles. Our analysis of user-reported relevance perception showed that the linear profile is a closer fit to the relevance perceptions of the "average" user. However, the distribution of magnitudes again suggests that attempting to fit a single profile (or view of relevance) for system evaluation is unlikely to be sufficient, if the expectation is that the evaluation should be reflective of a broad user base.

While on average our magnitudes were broadly equivalent to previous ordinal scales, the outstanding feature of our data was the wide range of scores that participants chose to employ in the judging task. In particular, at least half of the participants chose gain values that are not consistent with currently used values. Section 7.2 shows that using judgments made on a wide scale leads to different system rankings than judgments collected on a narrow scale. Recall that these scales are not imposed on the judge, as they are in all previous relevance judgment tasks in the literature, but are chosen by the participants themselves.

This is another key contribution of this study: when a priori categorical scales are used for relevance judgment tasks, there is no possible way to capture variance in human perception of the scale of relevance. In turn, this limits our understanding of how gain should be set in DCG-like metrics, and hence our ability to accurately evaluate systems.

9.2. Limitations and Future Work

This study is the first to investigate document level relevance judgments using magnitude estimation, at scale. A much-cited benefit of magnitude estimation is that it is a ratio scaling technique. Indeed, workers were instructed to assign relevance scores as ratios to those assigned to previous documents. However, some documents in the collected data lead to ratios that have a large number of significant figures. For example, in one unit, the first document was given a score of 54, and the second 76, implying that the worker thought that the second document was 1.407 times more relevant than the first. This seems like an unusually high amount of precision for such a task, and it may be that the worker was choosing numbers without closely following the ratio requirement. Intuitively, one might think that it would be possible to distinguish the relevance of two documents at a ratio of one significant figure (for example, "2 times as relevant", or "100 times as relevant"); or maybe even two significant figures ("21 times as relevant", or "140 times as relevant"); but it seems very unlikely that this could be done to three significant figures or more without extensive training and effort. We can safely assume that our crowdsourced workers were not trained in relevance assessment, nor spent more than a minute or two on each document (mean time to assess 78 seconds, standard deviation 86 seconds).

If we assess each unit of work using a criterion that each relevance score must have a ratio with at least one previously assigned score that has at most two significant figures, then 4806 of the 7059 units match. If we are more strict, and insist on only one significant figure, then only 3591 of 7059 units match. It seems, therefore, that about half of the units are being completed without a strict adherence to the idea of assigning numbers following ratio relationships. We note, however, that this might not be a problem in practice because of approximation and balancing effects: the 1.407 figure above is likely not exact, but may be a good approximation of a value between 1 and 2. Moreover, relevance is not a strictly defined concept, and the magnitude estimation is intended to reflect human perceptions. Overall, the obtained magnitude estimation scores were shown to have high external consistency with other judging scales.

Throughout the paper we have assumed that topical relevance collected using magnitude estimation can be used directly as gain in DCG-like measures. While perhaps being more representative of user perceptions than other relevance scales, this still makes many assumptions about the user's search process which are probably untrue: for example, ignoring the interdependence of documents and other aspects of relevance. In future work, we plan to investigate whether magnitude estimation can be used to reliably scale other aspects of relevance such as novelty, as well as search outcome

measures such as satisfaction, and to examine whether this can lead to more meaningful gain representations.

In this paper we focused on quantitative analysis of the magnitude estimation process. As part of the judging interface, judges were also asked to enter short textual comments to justify their choice of score. We plan to analyze this qualitative data in future work.

The bulk of the analysis in this paper considered magnitude estimation judgments as a reflection of relevance at the aggregate level. A further interesting line of analysis would be to consider the technique at the topic level, as different perceptions of relevance may be useful of indications of topic difficulty, or as potential calibration functions of how well a system could be expected to perform.

ACKNOWLEDGMENTS

This work was supported in part by the Australian Research Council (*Discovery Project* DP130104007) and in part by a Google Faculty Research Award. Thanks to both the SIGIR 2015 and TOIS reviewers who helped improve this manuscript.

REFERENCES

Omar Alonso and Stefano Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *Information Processing and Management* 48, 6 (2012), 1053–1066.

Jesse Anderton, Maryam Bashir, Virgil Pavlu, and Javed A. Aslam. 2013. An Analysis of Crowd Workers Mistakes for Specific and Complex Relevance Assessment Task. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. 1873–1876.

Lora Aroyo and Chris Welty. 2015. Truth Is a Lie: Crowd Truth and the Seven Myths of Human Annotation. *AI Magazine* 36, 1 (2015), 15–24.

Ellen Gurman Bard, Dan Robertson, and Antonella Sorace. 1996. Magnitude estimation of linguistic acceptability. *Language* 72, 1 (1996), 32–68.

Chris Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Greg Hullender. 2005. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*. 89–96.

Ben Carterette and Ian Soboroff. 2010. The effect of assessor error on IR system evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. 539–546.

Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. 621–630.

Charles Clarke, Nick Craswell, and Ian Soboroff. 2005. Overview of the TREC 2004 Terabyte Track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*. NIST.

Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2013. Examining the Limits of Crowdsourcing for Relevance Assessment. *IEEE Internet Computing* 17, 4 (July 2013), 32–38.

Kevyn Collins-Thompson, Paul Bennett, Fernando Diaz, Charles LA Clarke, and Ellen M Voorhees. 2014. TREC 2013 Web Track Overview. In *22nd Text REtrieval Conference (TREC 2013)*.

Eli P Cox. 1980. The optimal number of response alternatives for a scale: A review. *Journal of marketing research* 17, 4 (1980), 407–422.

Walter H. Ehrenstein and Addie Ehrenstein. 1999. Psychophysical Methods. In *Modern techniques in neuroscience research*, Uwe Windhorst and Hakan Johansson (Eds.). Springer, 1211–1241.

Michael Eisenberg. 1988. Measuring Relevance Judgements. *Information Processing and Management* 24, 4 (1988), 373–389.

George Gescheider. 1997. *Psychophysics: The Fundamentals* (3rd ed.). Lawrence Erlbaum Associates.

Andrew F Hayes and Klaus Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. *Communication methods and measures* 1, 1 (2007), 77–89.

Mehdi Hosseini, Ingemar J. Cox, Nataša Milić-Frayling, Gabriella Kazai, and Vishwa Vinay. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In *Proceedings of the European Conference on Information Retrieval*. 182–194.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

Korey Johnson and Aga Bojko. 2010. How Much is too Much? Using Magnitude Estimation for User Experience Research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 54, 12 (2010), 942–946.

HJ Jung and Matthew Lease. 2011. Improving Consensus Accuracy via Z-score and Weighted Voting. In *Proceedings of the 3rd Human Computation Workshop*. 88–90.

Evangelos Kanoulas and Javed A Aslam. 2009. Empirical justification of the gain and discount function for nDCG. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*. 611–620.

Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. 2013. An Analysis of Human Factors and Label Accuracy in Crowdsourcing Relevance Judgments. *Information Retrieval* 16, 2 (April 2013), 138–178.

Andrew Keen. 2008. *The Cult of the Amateur: How Blogs, MySpace, YouTube, and the rest of today's user-generated media are killing our culture and economy*. Random House Inc.

Klaus Krippendorff. 2011. Computing Krippendorff's alpha reliability. *Departmental Papers (ASC)* 1-25-2011 (2011).

Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2015. Judging Relevance Using Magnitude Estimation. In *Proceedings of the European Conference on Information Retrieval*. 215–220.

Lawrence Marks. 1974. *Sensory Processes: The new Psychophysics*. Academic Press.

Richard McCreadie, Craig Macdonald, and Iadh Ounis. 2011. Crowdsourcing blog track top news judgments at TREC. In *Proceedings of the ACM WSDM International Conference on Web Search and Data Mining*. 23–26.

Mick McGee. 2003. Usability magnitude estimation. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* 47, 4 (2003), 691–695.

Howard R. Moskowitz. 1977. Magnitude estimation: notes on what, how, when, and why to use it. *Journal of Food Quality* 1, 3 (1977), 195–227.

Tetsuya Sakai. 2007. On the reliability of information retrieval metrics based on graded relevance. *Information Processing and Management* 43, 2 (2007), 531 – 548.

Tefko 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. *Journal of the Association for Information Science and Technology* 58, 13 (2007), 1915–1933.

Falk Scholer, Eddy Maddalena, Stefano Mizzaro, and Andrew Turpin. 2014. Magnitudes of Relevance: Relevance Judgements, Magnitude Estimation, and Crowdsourcing. In *The Sixth International Workshop on Evaluating Information Access (EVIA 2014)*.

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

David Sheskin. 2007. *Handbook of parametric and nonparametric statistical procedures, 4th ed.* CRC Press.

Mark Smucker, Gabriella Kazai, and Matthew Lease. 2014. Overview of the TREC 2013 Crowdsourcing Track. In *Proceedings of the 22nd NIST Text Retrieval Conference (TREC)*.

Eero Sormunen. 2002. Liberal Relevance Criteria of TREC: Counting on Negligible Documents?. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. 324–330.

Amanda Spink and Howard Greisdorf. 2001. Regions and levels: Measuring and mapping users' relevance judgments. *Journal of the Association for Information Science and Technology* 52, 2 (2001), 161–173.

Stanley S Stevens. 1966. A metric for the social consensus. *Science (New York, NY)* 151, 3710 (1966), 530–541.

Rong Tang, William M Shaw, and Jack L Vevea. 1999. Towards the identification of the optimal number of relevance categories. *Journal of the American Society for Information Science* 50, 3 (1999), 254–264.

Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. ACM, 565–574.

Ellen M. Voorhees and Donna K. Harman. 1999. Overview of the Eighth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*. 1–24.

Ellen M. Voorhees and Donna K. Harman. 2005. *TREC: experiment and evaluation in information retrieval*. MIT Press.

William Webber and Jeremy Pickens. 2013. Assessor Disagreement and Text Classifier Accuracy. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. 929–932.

Yinglong Zhang, Jin Zhang, Matthew Lease, and Jacek Gwizdka. 2014. Multidimensional Relevance Modeling via Psychometrics and Crowdsourcing. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*. ACM, 435–444.

Guido Zuccon, Teerapong Leelanupab, Stewart Whiting, Emine Yilmaz, Joemon M Jose, and Leif Azzopardi. 2013. Crowdsourcing interactions: using crowdsourcing for evaluating interactive information retrieval systems. *Information retrieval* 16, 2 (2013), 267–305.

Appendix A: Instructions Shown to Participants

Introduction

In this task, you will be asked to make Magnitude Estimation assessments. Magnitude Estimation is a technique whereby you assign numbers to indicate your perception of the strength of an effect. It is described in more detail below.

Please read these instructions carefully before deciding whether to complete the task. Note that there are some checks throughout the task, and if you do not perform these correctly you will not be able to terminate and get paid. The data from this task is being gathered for research purposes. No personally identifying information is recorded. Participation is entirely voluntary, and you are free to discontinue at any point.

Task Steps

1. Magnitude Estimation of line lengths

As a warm-up task, to familiarize you with Magnitude Estimation, you will be shown a sequence of three lines, one at a time.

1.a For each line, enter a number into the text box below the displayed line. This number should reflect your perception of the length of the line. You may use any numbers that seem appropriate to you – whole numbers, fractions, or decimals. However, you may not use negative numbers, or zero.

1.b For each subsequent line, enter a number that reflects your perception of its length, relative to the previous line. For example, if you feel that the current line is twice as long as the previous, then you should assign a number that is twice as large as the number you used previously.

Don't worry about running out of numbers – there will always be a larger number than the largest you use, and a smaller number than the smallest you use.

Note: the Magnitude Estimation scores are *not* intended to be an estimate of the length in any particular measurement units, such as centimetres.

2. Magnitude Estimation of document relevance

In the main part of this task, you will be asked to assign Magnitude Estimation scores to indicate your perception of document relevance.

2.a First, a statement that expresses a need for information will be displayed at the top of the screen. Please read the statement carefully. You will be asked a question about the statement, to test your understanding.

2.b You will then be asked to rate 8 documents that have been returned by a search system, in response to the information need statement.

Your task is to indicate how RELEVANT these documents appear to you, in relation to the information need.

As a preliminary exercise, can you imagine a document which would be highly relevant to the information statement? Can you imagine a document that you would judge to be low in relevance? Can you imagine a document that you would judge to be medium in relevance?

Now do the same for numbers. Imagine of a large number. A small number. A medium number.

As indicated above, you will be shown 8 documents, one at a time. Your task will be to assign a number to every document in such a way that your impression of how large the number is matches your judgment of how relevant the document is.

Write the number for each document in the box under the document description.

— You may use any numbers that seem appropriate to you – whole numbers, fractions, or decimals. However, you may not use negative numbers, or zero.
— Don't worry about running out of numbers – there will always be a larger number than the largest you use, and a smaller number than the smallest you use.
— Try to judge each document in relation to the previous one. For example, if the current document seems half as relevant as the previous one, then assign a score that is half of your previously assigned score.
— You are requested to indicate your best judgment of relevance of a document at the time it is presented to you, one document at a time. However, if you wish to correct a mistake, then you can use the back button to revisit a previous judgement.
— The documents are not presented in any particular order. You might see many good documents, many bad documents, or any combination. Try not to anticipate, and simply rate each document after reading it.
— To judge each document, you will need to read it completely. A document's relevance (or non-relevance) might depend on a small part of the document. Don't try to guess, as there are some cross-checks, as indicated above.