The Benefits of Magnitude Estimation Relevance Assessments for Information Retrieval Evaluation

Andrew Turpin
University of Melbourne, Australia
aturpin@unimelb.edu.au

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

Stefano Mizzaro
University of Udine, Italy
mizzaro@uniud.it

Eddy Maddalena
University of Udine, Italy
eddy.maddalena@uniud.it

ABSTRACT

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 in the context of information retrieval evaluation, carrying out a large-scale user study across 18 TREC topics and collecting more than 50,000 magnitude estimation judgments. Our analysis shows that on average magnitude estimation judgments are rank-aligned with ordinal judgments made by expert relevance assessors. An advantage of magnitude estimation is that users can chose their own scale for judgments, allowing deeper investigations of user perceptions than when categorical scales are used.

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, in terms of how impactful relative differences in document relevance are perceived to be. We further use magnitude estimation to investigate gain profiles, comparing the currently assumed linear and exponential approaches with actual user-reported relevance perceptions. This indicates that the currently used exponential gain profiles in nDCG and ERR are mismatched with an average user, but perhaps more importantly that individual perceptions are highly variable. 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. Finally, we demonstrate that magnitude estimation judgments can be reliably collected using crowdsourcing, and are competitive in terms of assessor cost.

SIGIR '15, August 09–13, 2015, Santiago, Chile

ACM 978-1-4503-3621-5/15/08 … $15.00.
DOI: http://dx.doi.org/10.1145/2766462.2767760.

Categories and Subject Descriptors

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

Keywords

Magnitude estimation; evaluation; relevance assessments; relevance

General Terms

Measurement, performance, experimentation

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 [28]. 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 [26].

Magnitude estimation is 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 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 [18]. 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 [17].

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 [10], and our own small-scale pilot study [15, 21]. 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. Specifically, we consider three 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. 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?

RQ3. Can magnitude estimation relevance scores provide additional insight into user perceptions of relevance, into actual gain values, and into individual gain profiles?

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 analysis and discussion of our results, corresponding to the three research questions, is presented in Sections 4 to 6. Our conclusions and directions for future work are included in the final section of the paper.

2. BACKGROUND

Relevance: Relevance is an important concept in information science and information retrieval [19]. 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 [28]. 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 [19]. 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 reflect 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* [28]. 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 [12]. 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) [12] and *expected reciprocal rank* (ERR) [5]. Some examples of previous choices for the number of levels in ordinal relevance scales include 3 (TREC Terabyte Track [6]), 4 (TREC newswire re-assessments by Sormunen [23]), and 6 (TREC Web Track [7]). Tang et al. [26] 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 [8] concludes that there is no single number of response alternatives that is appropriate for all situations.

Magnitude estimation: Magnitude estimation 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 [9]. A key advantage of using magnitude estimation is that the responses are on a ratio scale of measurement [11], 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 [22].

Proposed by Stanley Stevens in the 1950s, magnitude estimation has a long history, and is the most widely-used psychophysical ratio scaling technique [11]. 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 [25]), in medical applications (such as levels of pain, severity of mental disorders, and emotional stress from life events [11]), in user experience research (for example, as a measure of usability in HCI [17], and for healthcare applications [13]), and in linguistics (including judging the grammaticality of sentences [2]).

In information retrieval research, Eisenberg [10] 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 [24], 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 [15, 21]. That study indicated that magnitude estimation could be useful for measuring document-level relevance, but used only 3 information need statements and 33 documents.

3. EXPERIMENTAL METHODOLOGY

In this section we describe the details of the experiments that were carried out to gather the needed relevance judgments.

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

Topic	# docs	# units	H_k	N_k	Back 0	Back 1	Back 2^+	Fail 0	Fail 1	Fail 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

Table 1: Topics, number of documents, number of units, H_k, N_k, number of back buttons usage, number of failures in the (c) and (d) checks, as detailed in the text.

NIST assessors [27]. Subsequently, Sormunen [23] 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 are the subset of TREC-8 topics which also have Sormunen ordinal judgments available. They are listed in the first column of Table 1. The *documents* for which we collected magnitude estimation judgments are 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 1 (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 1 (3rd column).

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.

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: *Your task is to indicate how relevant these documents appear to you, 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 1, 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 reassess the same topic.

Magnitude estimation assignments: Regarding the assignment of ME scores, participants were instructed that: *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.* (The complete instructions that were shown to participants are available at `http://cs.rmit.edu.au/~fscholer/ME/SIGIR15`.) 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 [18]; we therefore allowed participants to freely choose their own values.

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 [18]. 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 H_k and 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 1). 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 [11], and also has an impact on the score normalization process, described below. Moreover, including the two N_k and 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.

Quality checks: Three 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; and

(c) a check that the magnitude estimations score for $\mathtt{H}_k$ was greater than that assigned to $\mathtt{N}_k$.

In addition, a time-based quality check was also used:

(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 1, 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. As such, we did not do any further filtering once the data had been collected.

Crowd Judging: As detailed in Table 1, with 7,059 units comprised of 8 documents each, 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 [1, 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.

3.3 Score Normalization

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 [11, 17], and was applied in our data analysis.

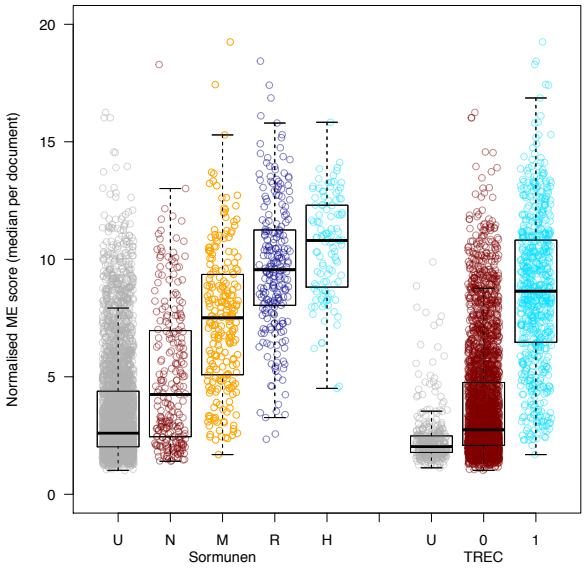

Figure 1: ME score distribution by Sormunen and TREC levels.

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 is taken. Next, the arithmetic mean of these log scores is calculated for each unit (μ_u), and for the topic as a whole (μ). 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:

$$s_i' = \exp(\log(s_i) - \mu_u + \mu)$$

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. 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 [16]. Importantly, normalization through geometric averaging has the property of preserving the ratios of the original scores, the essential feature of the magnitude estimation process. Unless otherwise noted, all analysis reported in this paper uses normalized magnitude estimation scores.

4. 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, 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.

4.1 Consistency of Magnitude Estimation and Ordinal Relevance

The distribution of the median normalized magnitude estimation scores for each document are shown in Figure 1, aggregated across all 18 topics, and split by Sormunen ordinal relevance levels (left side of figure, levels are $\mathtt{N}$, $\mathtt{M}$, $\mathtt{R}$, $\mathtt{H}$, and $\mathtt{U}$, the group of docu-

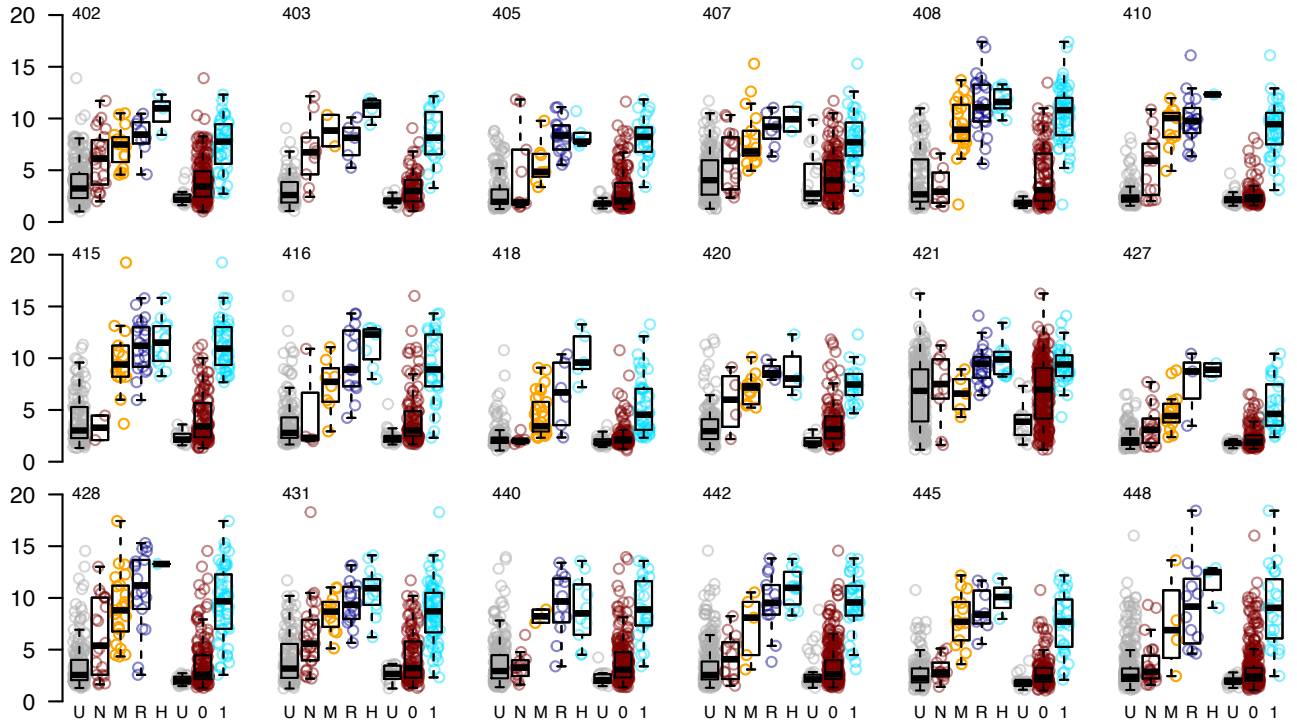

Figure 2: ME score distribution by Sormunen (U, N, M, R, H) and TREC (U, 0, 1) levels: breakdown for individual topics.

ments 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). This boxplot (and subsequent boxplots) show the median as a solid black line; boxes show the 25th to 75th percentile; and whiskers show the range, up to 1.5 times the inter-quartile range. 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 2 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 usually 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 some exceptions for some of the Sormunen adjacent categories in topics 403, 405, 410, 420, 421, and 440; there are no 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 1 and sub-plots in Figure 2), 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. It can also be observed that 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. We therefore next investigate judge agreement.

4.2 Judge Agreement

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. To investigate the level of agreement when using different scales for judging relevance, we compare the pairwise orderings between binary, ordinal and magnitude estimation relevance judgments. Figure 3 shows the proportion of document pairs that agree on the order of relevance. For example, when comparing ordinal and magnitude estimation ratings, if two documents are rated N and M, and the corresponding magnitude estimation score of the first is lower than or equal to the score for the second, this would be an agreement, while if the magnitude estimation score was higher for the second document, it would indicate disagreement. The figure shows the pairwise agreement proportions between magnitude estimation (ME), Sormunen (S) and TREC (T), for each of the six possible combinations of ordinal relevance levels. Red circles indicate the mean score 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 leads to a higher average agree-

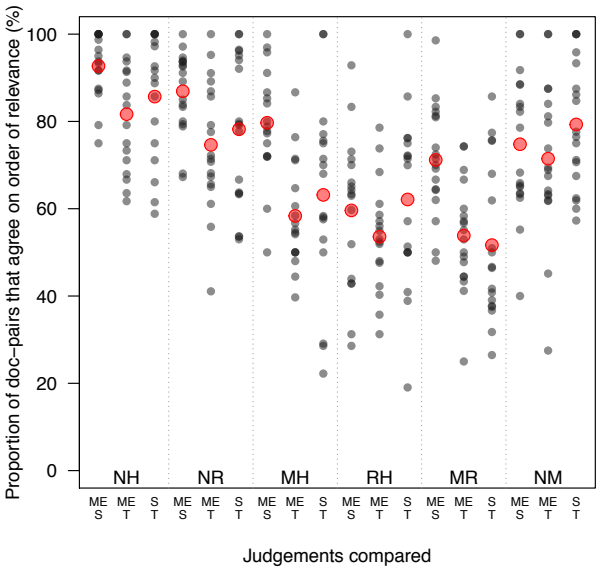

Figure 3: Agreement on the ordering of relevance of all document pairs (one small dot per topic) between judges as indicated on the x-axis (ME: magnitude estimation, S: Sormunen and T: TREC). The large (red) dot is the mean over all topics.

ment with Sormunen than TREC with Sormunen for documents in the NH, NR, MH, and MR pairs, and slightly lower for RH and NM: it can be said that normalized ME scores agree, in terms of ranks, with Sormunen at least as well as TREC agrees with Sormunen. Across all groups, the overall average agreement rates are 77% between magnitude estimation and Sormunen, 64% between magnitude estimation and TREC, and 65% between Sormunen and TREC. This further supports the validity of the magnitude estimation approach for gathering relevance judgments.

4.3 Failure Analysis

Despite the similar overall levels of agreement between the magnitude estimation method and ordinal relevance, Figures 1 and 2 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%) of 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%) of 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 [4, 20]; or they may display systematic errors due to a misunderstanding of the relevance criteria, or relevance drift [29]. 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.

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, having similar agreement rates with the ordinal Sormunen judgments as the Sormunen judgments have with TREC assessments.

5. MAGNITUDE ESTIMATION FOR SYSTEM-LEVEL EVALUATION

The second research question 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.

5.1 Gain in the nDCG and ERR Metrics

Magnitude estimation provides scores that reflect ratios of human perceptions of the intensity of relevance of different docu-

ments 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 [12] 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 [12]. 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 [5], 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 $R(G) = (2^{G_i} - 1)/2^{G_{max}}$. 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 [14] or 4 [12] 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.

5.2 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. Figure 4 shows the system scores obtained using median magnitude estimation scores per document (x-axis), and binary TREC categories (y-axis). It can be seen that there are definite changes in system rankings when using the different relevance scales. When using the nDCG@10 metric, there is only small movement of the top ranked systems (cluster of points towards the top right), but when using the ERR@10 metric there is a large perturbation of the top ranked systems using the different judgments.

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 $p < 0.05$. The overlap of the top set, based on the TREC and magnitude estimation relevance judgments is 44% for nDCG@10 and 76% 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 5 shows that there is also a large discordance between the system rankings using the Sormunen categories and median magnitudes as gain values. However, note that the scale is different in this figure than in Figure 4, as the simulated systems all have relevant documents in the top 10, and so are high scoring. The overlaps in top sets are 51% for nDCG@10 and 81% 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 both metrics, however.

5.3 Judgment Variability

There is substantial variation in the magnitudes assigned to documents by our judges. As we have at least 10 judgments per topic-document pair, rather than using the median score per document as in the previous section, we can resample individual scores many times, recompute system orderings, and get a distribution of τ values. The results are shown in Figure 6: the correlations between system orderings using sampled magnitudes as gains are generally lower than when using the median magnitudes. This reflects 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 1 it is apparent that magnitudes are generally in the range of 2 to 15 (although some are as large as

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

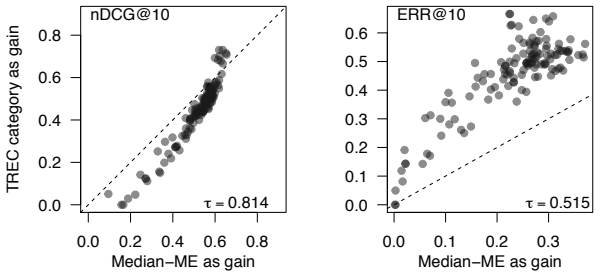

Figure 4: System scores using TREC categories (y-axis) and magnitudes (x-axis) as gains in the nDCG@10 and ERR@10 metrics. There is one dot per system participating in the TREC-8 ad hoc track. Kendall's τ is shown.

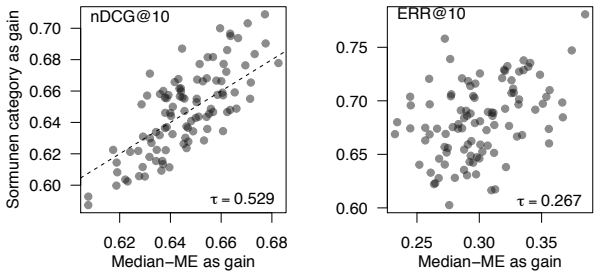

Figure 5: System scores using Sormunen categories (y-axis) and magnitudes (x-axis) as gains in the nDCG@10 and ERR@10 metrics. There is one dot per simulated system (see text). Kendall's τ is shown.

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 2 shows this for our 12 topics on the simulated systems. The final row hints that using magnitudes that are gathered on a wide scale may alter system rankings using nDCG@10 compared with simply using the category values as gains. Similarly, the first row suggests that using gains gathered on a wide scale may alter rankings using ERR@10. 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

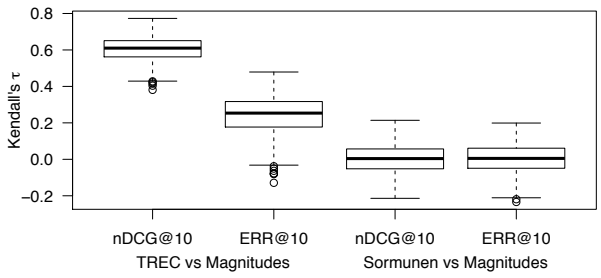

Figure 6: Kendall's τ between system rankings obtained using gains from the judgments and metrics indicated on the x-axis, where magnitudes are randomly sampled for each document 1000 times. Compare with τ in Figures 4 and 5.

Mapping	nDCG@10	ERR@10
$G_i = 100 \times C_i$	1.0000	0.8719
$G_i = 1 + C_i$	0.9994	0.9985
$G_i = 10 + C_i$	0.9950	0.9949
$G_i = 2^{1+C_i}$	0.9568	0.8738
$G_i = 10^{1+C_i}$	0.7833	0.8719

Table 2: Correlation (Kendall's τ) between system orderings when using Sormunen categories as gain ($G_i = C_i$), and other mappings on the simulated systems, when using the nDCG@10 and ERR@10 metrics.

Source of gains	Narrow	All	Wide
TREC, τ			
nDCG@10	0.780	0.814	0.765
ERR@10	0.537	0.515	0.434
TREC, top set			
nDCG@10	89%	44%	100%
ERR@10	56%	76%	95%
Sormunen-Simulated, τ			
nDCG@10	0.491	0.529	0.361
ERR@10	0.179	0.267	-0.010
Sormunen-Simulated, top set			
nDCG@10	67%	51%	82%
ERR@10	100%	81%	100%

Table 3: 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.

than 5 (the median value), and *wide* units, where the ratios $\mathtt{H}_k/\mathtt{N}_k$ are 5 or more.

Table 3 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 4 and 5). 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 minimum or maximum magnitude over all units was taken, respectively. System orderings based on narrow or wide units show greater differences (lower τ values) than when considering the full data set. This is consistent with the results in Table 2. The overlap of the top set is generally less than 100% for the narrow/wide data sets, indicating that some perturbation of the set of "equivalent" top systems is occurring, but the pattern is not as clear as for τ values. This may be due to the relatively small number of topics that is being considered, especially for the simulated systems, and more topics are required before a definitive conclusion about the effect of narrow and wide units on the top set can be made.

Overall, there are clear differences in system effectiveness orderings as measured using Kendall's τ, when using narrow and wide units. In particular, it appears that using the median of narrow units, or of all units, leads to higher correlations with existing measurement techniques, compared to when using wide units. We

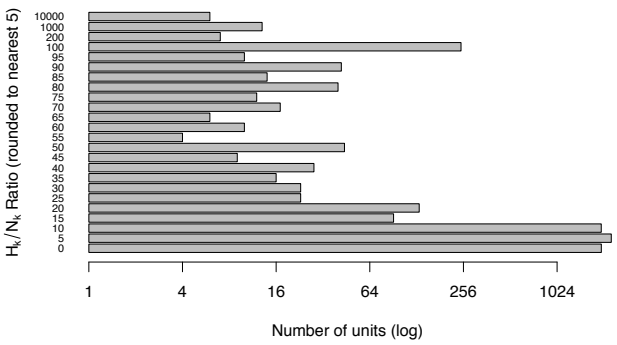

Figure 7: Number of units with a particular $\mathrm{H}_k/\mathrm{N}_k$ ratio over all topics. There were a total of 7,059 units. 37 units with frequency less than 4 are not shown.

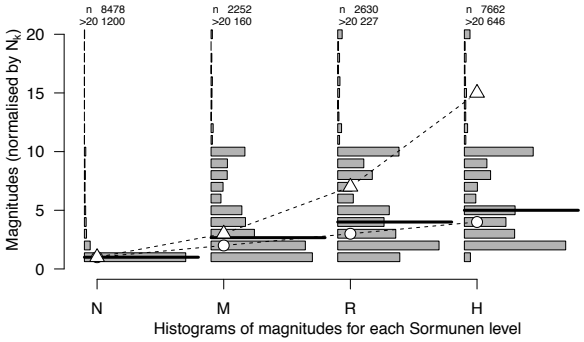

Figure 8: Distribution of magnitudes (normalized by N_k) for each Sormunen level. Circles are linear gain values, $\{1, 2, 3, 4\}$, while triangles represent exponential gain values, $\{2^1, 2^2, 2^3, 2^4\}$. Black lines are medians; the text "n" gives the total number of magnitudes in the level, and "> 20" shows the count of scores above 20.

can therefore infer that current evaluations using nDCG@10 and ERR@10 do not assume wide units as their underlying user model. This is important because it shows that nearly half of our participants are not behaving according to the user model assumed by current evaluation techniques. Figure 7 shows the full distribution of the "width" ($\mathrm{H}_k/\mathrm{N}_k$ ratio) of units rounded to the nearest 5 (with bars containing less than 4 units omitted).

6. 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 third research question 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 5); or an *exponential* setting, where the ordinal relevance level forms a power of 2 [3], placing more emphasis on highly relevant documents. Figure 8 shows the distribution of magnitude estimation scores, normalized by the 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.

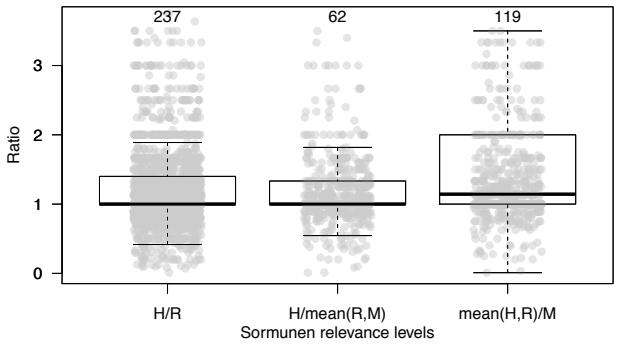

Figure 9: Ratio of the magnitude scores assigned to the Sormunen categories highly relevant (H) to relevant (R) and marginally relevant (M) over all pairs of documents. Numbers indicate the count of document pairs that are greater than 3.5 for each box.

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.

Our final piece of analysis is given in Figure 9, which shows the ratio of magnitudes for document pairs in the Sormunen levels highly relevant and partially relevant. We combine the three relevance levels into two to allow comparison with Kanoulas and Aslam [14]. They recommend a ratio of 1.1 for relevant to highly relevant, which is supported by the medians in Figure 9. However, again, we see a wide variation from this median.

7. DISCUSSION AND CONCLUSIONS

This is the first study to have collected large scale, real user data about relevance perceptions using magnitude estimation. Our first research question investigated the suitability of using magnitude estimation for gathering relevance judgments. As shown in Section 4, magnitude estimation scores are consistent with classical ordinal scores, both on single topics and overall on the 18 aggregated topics; ordering agreement between magnitude estimation judgments and ordinal judgments is higher than between ordinal and binary judgments; and failure analysis demonstrated that disagreement is often due not to problems with magnitude estimation per se but rather to different judging context. It is likely that using a classical scale one would get a similar agreement level.

Considering the impact of magnitude estimation relevance judgments on IR system evaluation, the second research question, our analysis showed that system orderings vary substantially when different relevance scales are used, with a τ of 0.81 for nDCG@10 when comparing magnitude estimation scores with binary relevance across TREC runs, and 0.53 when comparing with ordinal relevance on simulated runs. When individual variability is considered, these correlations are even lower. The analysis suggests that

it is important to incorporate different judging scales – users who have wider or thinner perceptions of the differences in relevance.

We also employed magnitude estimation to directly investigate gain profiles, the third research question. Typically, gains are set as linear or exponential profiles. Our analysis of user-reported relevance perception showed that the linear profile is close to the "average" user, but the distribution of magnitudes suggests that attempting to fit a single profile, or view of relevance, for system evaluation is unlikely to be sufficient.

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 5.2 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.

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.

Acknowledgments This work was supported by the Australian Research Council (*Discovery Project* DP130104007).

References

[1] O. Alonso and S. Mizzaro. Using crowdsourcing for TREC relevance assessment. *Information Processing and Management*, 48(6):1053–1066, 2012.

[2] E. G. Bard, D. Robertson, and A. Sorace. Magnitude estimation of linguistic acceptability. *Language*, 72(1):32–68, 1996.

[3] C. Burges, T. Shaked, E. Renshaw, A. Lazier, M. Deeds, N. Hamilton, and G. Hullender. Learning to rank using gradient descent. In *Proceedings of the 22nd international conference on Machine learning*, pages 89–96, Bonn, Germany, 2005.

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

[5] O. Chapelle, D. Metlzer, Y. Zhang, and P. Grinspan. Expected reciprocal rank for graded relevance. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, pages 621–630, Hong Kong, 2009.

[6] C. Clarke, N. Craswell, and I. Soboroff. Overview of the TREC 2004 terabyte track. In *The Fourteenth Text REtrieval Conference (TREC 2005)*, Gaithersburg, MD, 2005. NIST.

[7] K. Collins-Thompson, P. Bennett, F. Diaz, C. L. Clarke, and E. M. Voorhees. TREC 2013 Web Track Overview. In *22nd Text REtrieval Conference (TREC 2013)*, Gaithersburg, MD, 2014.

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

[9] W. H. Ehrenstein and A. Ehrenstein. Psychophysical methods. In U. Windhorst and H. Johansson, editors, *Modern techniques in neuroscience research*, pages 1211–1241. Springer, 1999.

[10] M. Eisenberg. Measuring relevance judgements. *Information Processing and Management*, 24:373–389, 1988.

[11] G. Gescheider. *Psychophysics: The Fundamentals*. Lawrence Erlbaum Associates, 3rd edition, 1997.

[12] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Transactions on Information Systems*, 20(4):422–446, 2002.

[13] K. Johnson and A. Bojko. How much is too much? Using magnitude estimation for user experience research. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 54(12):942–946, 2010.

[14] E. Kanoulas and J. A. Aslam. Empirical justification of the gain and discount function for nDCG. In *Proceedings of the ACM CIKM International Conference on Information and Knowledge Management*, pages 611–620, Hong Kong, 2009.

[15] E. Maddalena, S. Mizzaro, F. Scholer, and A. Turpin. Judging relevance using magnitude estimation. In *Advances in Information Retrieval - Proceedings of 37th ECIR*, volume 9022 of *LNCS*, pages 215–220, Vienna, Austria, 2015.

[16] L. Marks. *Sensory Processes: The new Psychophysics*. Academic Press, 1974.

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

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

[19] T. Saracevic. Relevance: A review of the literature and a framework for thinking on the notion in information science. Part II: nature and manifestations of relevance. *JASIST*, 58(13):1915–1933, 2007.

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

[21] F. Scholer, E. Maddalena, S. Mizzaro, and A. Turpin. Magnitudes of relevance: Relevance judgements, magnitude estimation, and crowdsourcing. In *The Sixth International Workshop on Evaluating Information Access (EVIA 2014)*, Tokyo, Japan, 2014.

[22] D. Sheskin. *Handbook of parametric and nonparametric statistical procedures, 4th ed.* CRC Press, 2007.

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

[24] A. Spink and H. Greisdorf. Regions and levels: Measuring and mapping users' relevance judgments. *JASIST*, 52(2):161–173, 2001.

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

[26] R. Tang, W. M. Shaw, and J. L. Vevea. Towards the identification of the optimal number of relevance categories. *JASIS*, 50(3):254–264, 1999.

[27] E. M. Voorhees and D. K. Harman. Overview of the Eighth Text REtrieval Conference (TREC-8). In *The Eighth Text REtrieval Conference (TREC-8)*, pages 1–24, Gaithersburg, MD, 1999.

[28] E. M. Voorhees and D. K. Harman. *TREC: experiment and evaluation in information retrieval*. MIT Press, 2005.

[29] W. Webber and J. Pickens. Assessor disagreement and text classifier accuracy. In *Proceedings of the ACM SIGIR International Conference on Research and Development in Information Retrieval*, pages 929–932, Dublin, Ireland, 2013.