# On Fine-Grained Relevance Scales

Kevin Roitero
University of Udine
Udine, Italy
roitero.kevin@spes.uniud.it

Eddy Maddalena
University of Southampton
Southampton, U.K.
e.maddalena@soton.ac.uk

Gianluca Demartini
University of Queensland
Brisbane, Australia
g.demartini@uq.edu.au

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

## ABSTRACT

In Information Retrieval evaluation, the classical approach of adopting binary relevance judgments has been replaced by multi-level relevance judgments and by gain-based metrics leveraging such multi-level judgment scales. Recent work has also proposed and evaluated unbounded relevance scales by means of Magnitude Estimation (ME) and compared them with multi-level scales. While ME brings advantages like the ability for assessors to always judge the next document as having higher or lower relevance than any of the documents they have judged so far, it also comes with some drawbacks. For example, it is not a natural approach for human assessors to judge items as they are used to do on the Web (e.g., 5-star rating).

In this work, we propose and experimentally evaluate a bounded and fine-grained relevance scale having many of the advantages and dealing with some of the issues of ME. We collect relevance judgments over a 100-level relevance scale (S100) by means of a large-scale crowdsourcing experiment and compare the results with other relevance scales (binary, 4-level, and ME) showing the benefit of fine-grained scales over both coarse-grained and unbounded scales as well as highlighting some new results on ME.

Our results show that S100 maintains the flexibility of unbounded scales like ME in providing assessors with ample choice when judging document relevance (i.e., assessors can fit relevance judgments in between of previously given judgments). It also allows assessors to judge on a more familiar scale (e.g., on 10 levels) and to perform efficiently since the very first judging task.

## CCS CONCEPTS

• **Information systems** → **Information retrieval**; **Relevance assessment**;

## KEYWORDS

IR Evaluation, Relevance Scales

**ACM Reference Format:**
Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On Fine-Grained Relevance Scales. In *SIGIR '18: The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, July 8–12, 2018, Ann Arbor, MI, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3209978.3210052


*SIGIR '18, July 8–12, 2018, Ann Arbor, MI, USA*

ACM ISBN 978-1-4503-5657-2/18/07.
https://doi.org/10.1145/3209978.3210052

## 1 INTRODUCTION

Relevance assessment is an integral part of Information Retrieval (IR) evaluation, since the Cranfield experiments, through the TREC and TREC-like initiatives. In the recent years the collection of relevance judgments is being studied using crowdsourcing. To gather relevance labels, several scales have been used in the past. The most common are the classical binary scale, or ordered scales with a limited number of categories (usually ranging from 3 to 7). It has recently been proposed [15, 21] to use Magnitude Estimation (ME) to gather relevance assessments on a $]0, +\infty[$ scale that has the following advantages:

(1) it is more fine-grained than the above alternatives (and thus, at least potentially, allowing to capture relevance differences that would otherwise be lost);
(2) it is able of always providing to the assessor a smaller or higher relevance value, and even a value in between other two, always allowing to assign to a new document a relevance value unforeseen in advance. This happens in particular at the extremes of the scale, but also for the values internal to the range; and
(3) it can adapt to different assessors' preferences (e.g., those who prefer to use a binary scale can do that and those who prefer to judge in a scale from 1 to 10 can also do that).

ME is not free from disadvantages, though:

- it requires a normalization of the collected scores since each assessor is free to use a different "internal" relevance scale. This normalization is not simple, and it is not clear which is the best alternative, although some techniques seem to be reasonably effective [15, 21];
- it does not allow for a direct comparison of scores provided by different judges and/or on different topics as the score normalization is typically performed on a topic-by-topic basis;
- it is somehow unnatural, or at least it requires some adaptation for the human assessor as compared to most common rating scales which are bounded; and
- it leads to a log-normal distribution of relevance scores.

In this paper we discuss and experimentally evaluate by means of large-scale crowdsourced relevance judgments the use a fine-grained scale on 100 levels (S100). Using the proposed 100 levels scale, the human assessor judges the relevance of a document with respect to a query by means of a number in the [0..100] range (extremes included, thus the levels are actually 101; we name it S100 anyway). Such a scale can be seen as a sort of compromise between the classical a-few-categories relevance scales and ME. We run a large scale crowdsourcing experiment to collect more than 50 thousand labels on such a scale, we discuss its advantages and disadvantages with respect to the already proposed alternatives, and we experimentally compare our judgments with judgments on

coarse-grained scales (i.e., binary and 4-levels) and with judgments using ME.

More specifically, our research questions are:

- Can relevance values be collected in a reliable way using a 100 levels scale in a crowdsourcing setting?
  - How do crowd workers choose to use the proposed scale?
  - Are the collected labels consistent with standard ground truths?
- What are the differences between S100 and ME?
  - What are the effects of the relevance scale on IR system evaluation/ranking?
  - Which scale is more robust to decreasing the number of judgments per topic/document pair? What happens when collecting fewer judgments per assessor?
  - ME requires some learning to be used effectively by crowd workers as it is not like rating scales they are already used to. What happens to judgment quality when the number of documents judged by each worker in each HIT[1] decreases?
  - ME allows to go beyond the maximum and minimum judgment level previously used, and to always find a judgment in between two previously expressed judgments. Are these properties required and useful in practice when using S100?
  - Are S100 and ME different w.r.t. the time needed to express judgments? Does ME require more adaptation time when used for the first time (i.e., on the first documents)?

Our main findings are:

- w.r.t. binary and coarse-grained relevance scales, S100 gives assessors more flexibility in terms of preferential judgments over the documents they are presented during the judging task. Assessors using S100 also have the freedom to judge on a 10-level scale (or 4-level, etc.). It also better aligns with coarse-grained scales as compared to ME (see Section 4).
- w.r.t ME, S100 gives assessors a reference point by providing upper and lower scale boundaries (see Section 4).
- S100 is more robust than ME to both fewer assessors per document and fewer documents per assessor (see Section 5).
- The theoretical problem of running out of values (at the extremes of the scale) does not occur often in practice, at least in our setting. Of course, with more document to judge for each worker, the problem might manifest (see Section 6).
- If a fine-grained scale is preferred, using ME in a crowdsourcing setting can provide results faster while S100 enables direct comparison over topics and workers and does not require normalization (see Section 7).
- While ME shows a steeper learning curve with more time needed to judge the first few documents, it becomes faster for crowd workers to judge with ME compared to S100 in the long term. Considering that crowd work is long-tail distributed with most workers completing very few HITs, S100 may be a more efficient strategy for crowdsourced relevance judgments (see Section 7).

This paper is structured as follows. Section 2 surveys related work in the area of relevance scales and crowdsourced relevance judgments. Section 3 presents an analysis of the relevance judgments we collected on the newly proposed S100 scale. In Section 4 we compare S100 with other commonly used bounded scales with 2 and 4 levels and with the ME unbounded scale. Section 5 compares the robustness of S100 and ME to having few judgments per document and few documents per assessor. Section 6 presents an analysis looking at how S100 may lead to assessors running out of values as compared to ME which enables higher Section 7 compares S100 and ME in terms of the time required for assessors to express their judgments. Section 8 summarizes our main findings and the main benefits and drawbacks of the newly proposed S100 scale as compared to commonly used relevance scales.

[1]Human Intelligence Task, the task that each worker has to perform.

## 2 RELATED WORK

### 2.1 Binary vs Graded Relevance Judgments

Relevance is a central concept in IR [18] evaluation; IR systems are usually evaluated using test collections, which are composed of (i) a collection of documents, (ii) a set of queries (called topics), and (iii) a set of relevance assessment for each (topic, document) pair in a pooled set of documents; such assessments are made by human experts according to an ordinal scale, which is usually binary.

Test collections can be created by means of a competition: participating systems return a ranked list of $n$ documents (usually 1000), which are then used to compute the judging pool (e.g., the top 100 documents returned by each system, for each topic). The documents in the pool are the ones assessed by human experts. The produced relevance judgments are used together with the ranked lists produced by the systems to compute an effectiveness metric for each (system, topic) pair; a commonly used metric is Average Precision (AP). In order to provide a final rank of participant systems, the effectiveness scores are averaged over the set of topics; for example, the average of AP scores originates Mean AP (MAP).

Historically, relevance judgments were made by assessing whether a document is relevant or not to a topic; then, based on the observation that more than two levels might be needed, a set of novel metrics which incorporate multiple levels relevance scales were developed, such as, for example, Normalized Discounted Cumulative Gain (NDCG) [10], Expected Reciprocal Rank (ERR) [2], and Q-Measure [17].

Concerning the ideal number of relevance levels to be used, over the years many proposal have been made: a three-level scale was used in TREC-Terabyte Track [3], a six-level scale was used in TREC-Web Track [4], a seven-levels scale was proposed by Tang et al. [20] when studying evaluation of bibliographic records by students, using relevance scales with a range of levels from two to eleven. Then, Maddalena et al. [15] proposed an unbounded scale based on Magnitude Estimation, which is described in the following. Despite the many different approaches on relevance scales, the question of how many relevance levels should we use is far from answered. In our work we present a comprehensive study on the effects of relevance scales on IR evaluation proposing a fine-grained scale at 100 levels that incorporates the benefits of both bounded scales as well as the flexibility of an unbounded scale.

### 2.2 Continuous Relevance and Magnitude Estimation

We provide some more details on the use of ME since we compare against it in the following, and since our experiments rely on reassessing documents on a 100-level scale following the same

experimental setting. ME is psychophysical technique used to measure the intensity of sensations [15]. The ME technique asks a human subject to give as a first response a number in the range $(0; +\infty)$; the successive numbers are assigned to reflect their relative difference; the outcome of ME are a set of measurements in a ratio scale [7]. Maddalena et al. [15] evaluated, using the CrowdFlower[2] platform, 18 TREC-8 topics, for a total of 4,269 documents. The documents are the top 10 documents returned for IR systems competing in the ad-hoc track; some documents (i.e., 3,881) were previously evaluated by TREC assessors using a binary scale, and some of those documents (i.e., 805) have been reassessed in the study by Sormunen [19] using a 4-level scale.

Results from [15] show that: (i) ME aggregated judgments are closely aligned with the ordinal coarse-grained scale, both overall and across topics; (ii) the gathered judgments have shown a high level of agreement with both TREC and Sormunen; (iii) the impact on system evaluation, i.e., the correlation between the system ranking when using ME judgments, has a Kendall's $\tau$ correlation of 0.677 with the official TREC ranking using NDCG@10.

In this paper we look at the challenges and opportunities of using S100 as compared to ME, binary and 4-level scales by means of comparative experiments using crowdsourcing platforms to collect relevance judgments at scale.

## 2.3 Relevance Dimensions and Biases

Recently, Jiang et al. [12] looked at the use of a multi-dimensional relevance definition including novelty, understandability, reliability, and effort for contextual judgments that are performed by assessors when looking at the search engine result page. In our work we rather focus on the classic definition of relevance based on topicality and look at the effect of different scales on IR evaluation.

Eickhoff [5] looked at the effect of cognitive biases in crowdsourced relevance judgment tasks. He showed how crowd workers are affected by fellow workers' answer (Bandwagon effect) and by being presented with multiple options (Decoy effect). The existence of the Decoy effect proves that workers judgment is indeed affected by other documents they have seen before judging a given document, thus supporting even more the need for fine-grained relevance scales (as we propose in our work) that enable workers to express slight relevance differences across different documents.

## 2.4 Crowdsourcing for IR Evaluation

Over the last few years, the increasing size of document collections created the need to scale the gathering of relevance judgments. For this reason, crowdsourcing has become a consolidated methodology to create relevance labels for query-document pairs given a judgment pool. In order to produce crowdsourced relevance labels at a quality level comparable with that of expert assessors a number of techniques have been proposed and evaluated in literature. A common approach is to collect relevance judgments for the same query-document pair from different crowd workers and to aggregate them together [1, 8, 22] thus allowing to remove noise in the labels. Past research also showed that asking for a justification for the judgments [16] and that limiting the time to judge [14] can increase crowdsourced relevance judgment quality. In our work we leverage crowdsourcing to collect relevance judgments over different scales and build on top of existing crowdsourcing research in terms of quality checks and HIT design best practices.

[2]https://www.crowdflower.com/

# 3 S100: A 100-LEVEL RELEVANCE DATASET

In this section, we present the results of our crowdsourcing effort aimed at collecting judgments on a 100-level scale. To make our dataset comparable with others, we followed the experimental design defined by [15] and reassessed 4,269 documents from 18 topics of TREC-8 ad hoc collection, in a [0, 100] discrete scale. As done in [15], we used the CrowdFlower crowdsourcing platform and rewarded workers $0.2 for each HIT performed (defined as a sequence of 8 documents which required to be judged in relation to one topic).

The main design difference as compared to that used for the ME collection by [15] is in the HIT graphical interface which, in our case, expects the relevance score to be given by using a [0,100] slider, instead than using a text field and an unbounded scale. The adoption of the slider is motivated by the bounded and fine-grained scale and it commonly used for rating items on multi-level scales (see, for example, [9]). In terms of quality checks, we performed the same checks as [15] (i.e., a test question on topic understanding; at least 20 seconds spent on at least 6 of the 8 documents in the HIT; consistency of judgments on two gold documents included in the 8 documents). Additionally, we required workers to move the slider (which was pre-set at 50) for at least 4 of the 8 documents. When failing the quality checks, workers were allowed to restart the HIT and change their previous answers. Up to 3 attempts were allowed. We tracked the times spent by each worker on each document, and these were cumulated over different attempts. We observed that 85.3% of workers completed the HIT after the first attempt, 11.2% after the second, and 3.5% after the third. Workers could not work on a topic more than once, but they were given the chance to repeat the task on different topics.

## 3.1 Judgment Distribution in S100

Figure 1 (a) shows the distribution of the individual scores gathered for S100: the x-axis represent the score obtained by a document, the y-axis represent its frequency; the red line represent the cumulative distribution. Figures 1 (b) and (c) show the distribution when doing a breakdown on non-relevant documents and on relevant documents, according to TREC assessors, respectively. From the plot using all the judgments (Figure 1 (a)) we see that, as expected, the distribution is clearly skewed towards lower (less relevant) scores; furthermore, there is a clear tendency of giving scores which are a multiple of ten, and the two most frequent scores are 0 and 100. In fact, the scores which are divisible by 10 are 60% of all the judgments in the dataset. The judgments on the scale boundaries (i.e., 0 and 100) are the 41%. If we do not consider the scale boundaries, the number of judgments which are divisible by 10 are the 32%. Due to the large presence of non relevant documents, the total distribution of scores is mainly influenced by and very similar to the distribution of the non-relevant documents according to TREC assessors (Figure 1 (b)), as we can see when comparing the cumulative distribution for the plots of all the documents with the one of non-relevant documents (i.e, the red lines in Figure 1 (a) and (b)).

When comparing Figure 1 (b) and (c), we see that for non-relevant documents (Figure 1 (b)) the majority of S100 scores is in the lower

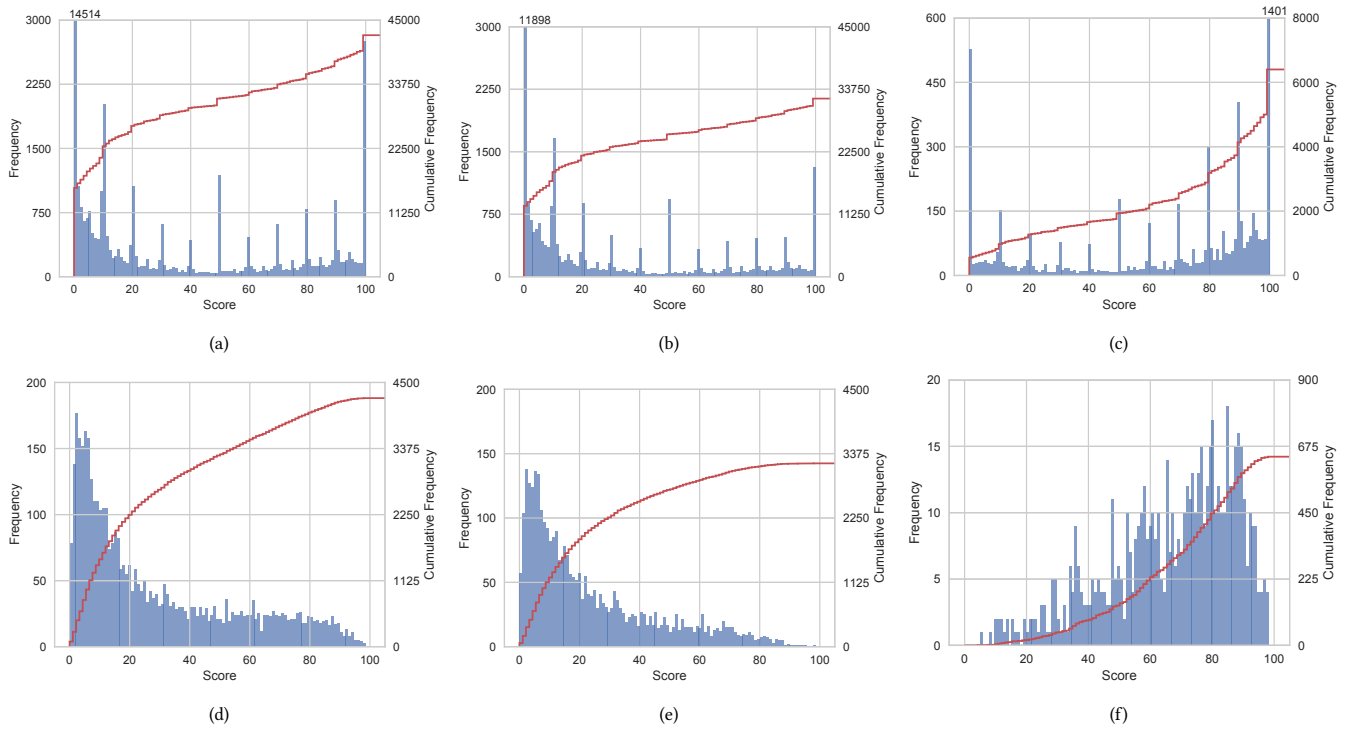


**Figure 1: Individual score distribution in the S100 dataset for all (a), for non-relevant (b), and for relevant (c) documents according to TREC judgments. Aggregated score distribution in the S100 dataset for all (d), for non-relevant (e), and for relevant (f) documents according to TREC judgments.**

part of the scale (left on the plot) and for relevant documents (Figure 1 (c)) the majority of S100 scores is in the higher part of the scale (right on the plot). Moreover, we observe that many non-relevant documents obtained the maximum possible score (i.e., 100), and many relevant documents obtained 0 as a score. This may depend on multiple factors: a misclassification by TREC experts, a document/topic ambiguity, or might even be an indicator of low quality crowd judgments, obtained despite the strict quality checks applied to the task. Furthermore, we notice that the "decimal preference" is still present both for relevant and non relevant documents.

## 3.2 Aggregated Judgments in S100

Next, we proceed with aggregating the raw relevance judgments collected form the crowd for the same topic/document pair as commonly done to increase the quality of the collection. Relevance scores in S100 are in the [0, 100] range, thus a natural aggregation function is represented by the arithmetic mean of the individual scores, with no prior normalization of individual scores as done for ME.[3] Figure 1 (d,e,f) show the distribution for the aggregated judgments. We can see that, as compared to the raw judgments given by individual workers (Figure 1 (a,b,c)), the aggregated judgments follow the expected distribution of many non-relevant documents with a long-tail of more relevant documents. The aggregation has also the effect of making the curves smoother, as well as making the tendency of scores to be a multiple of ten less prominent.

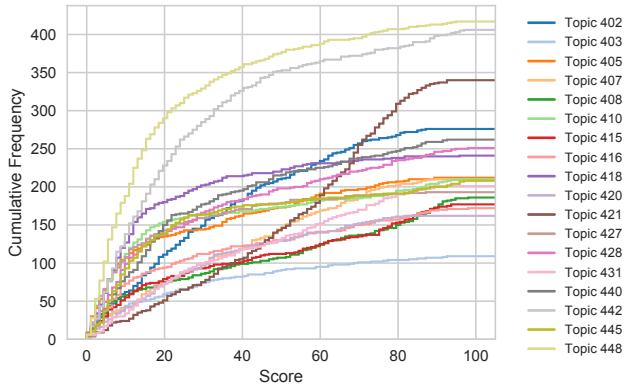


**Figure 2: Aggregated score cumulative distribution in the S100 dataset; breakdown on individual topics.**

Figure 2 shows the aggregated judgment cumulative distributions broken down by topic. We can observe similar trends over all topics with some topics (e.g., 448 and 442) having a cumulative curve growing faster (i.e., having many 'not so relevant' documents) and others having a much slower growth (e.g., 403) showing a presence of more relevant documents. A slightly different pattern is shown by topic 421 which grows towards the end of the relevance

[3] We experimentally compared different aggregation functions and the use of score normalization functions but observed that the use of the arithmetic mean over non-normalized score lead to most accurate labels compared to the other datasets.

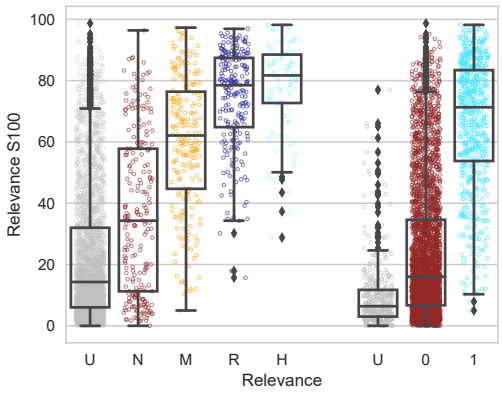


**Figure 3: Aggregated judgment scores collected for S100 vs TREC labels (right) and vs S4 (left). U indicates unjudged documents in the TREC and Sormunen collections.**

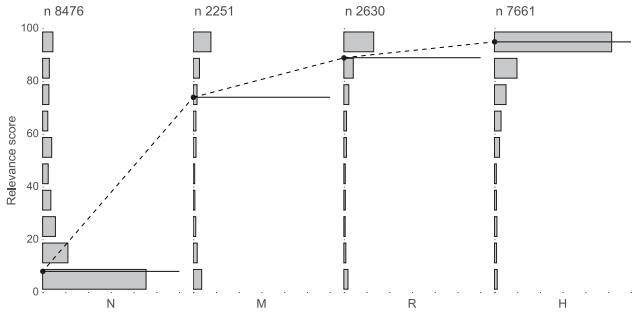


**Figure 4: Distribution of individual scores in S100 for each Sormunen level.**

score interval. This is explained by the fact that this topic has a small fraction of low relevance documents (Sormunen 0 and 1) and a high fraction of high relevant documents (Sormunen 2 and 3).

## 4 COMPARISON WITH OTHER SCALES

In this section we compare the judgments collected for S100 with judgments performed on the same documents over different relevance scales. We introduce an agreement measure that allows us to compute agreement across judgment scales and report agreement values for S100 with the TREC binary scale, the Sormunen 4-level scale (S4), and ME.

### 4.1 Score Distribution as Compared to Other Scales

Figure 3 shows how the judgments performed on S100 compared with the binary labels collected by TREC and the 4-level judgments performed by Sormunen. We can observe that while the median value for documents judged as relevant and non-relevant by TREC is different in S100, the distribution of S100 scores covers the entire score interval for both type of documents. The distribution of S100 scores compared to S4 labels by Sormunen shows the non linearity of the 4-level labels that have been collected on the scale N-M-R-H (i.e., not relevant, marginally relevant, relevant, and highly relevant). When comparing to the similar Figure 3 by Maddalena et al. [15] we can notice that in S100 the relevance scores are better distributed across the full scale with highest levels of relevance in S4 and the binary TREC scale having a median score closer to the upper bound of the scale as compared to ME. Such behavior is not observed for ME as the scale is unbounded at the top making scores for highly relevant documents having a wider distribution.

Figure 4, similar to Figure 14 by Maddalena et al. [15], presents a clearer evidence of the difference between ME and S100: whereas in ME the gain profiles seem to be exponential, or at least super-linear, in our case they are clearly sub-linear. We consider this a reason to prefer S100 to ME, since it better reflects the definition of relevance levels introduced for S4 which assumes that already marginally relevant documents should be substantially better than not relevant ones with a sub-linear step increase for the subsequent relevance levels R and H. Such differences (Figures 3 and 4) between ME and S100 scores are likely due to the effect of the end of scale which is unbounded in ME thus making high-relevance scores disperse. This is not the case in S100, a bounded scale that allows assessors to implicitly map their judgments against the scale upper bound.

### 4.2 Agreement with TREC

First we introduce a new measure that allows us to check assessor agreement across rating scales. We then adopt this measure to evaluate the quality of S100 scores as compared to other datasets.

*4.2.1 An agreement measure for ratings given over different scales.* Given two rating vectors $X = \{x_1, \ldots, x_n\}$ and $Y = \{y_1, \ldots, y_n\}$ where $x_i$ ($y_i$ respectively) represents the $i$-th document in a sequence of relevance judgments (e.g., a HIT), we define $X'$ as the sorted vector X and $Y'$ the re-ordering of $Y$ maintaining the relation to $X'$. That is, $X' = \{x_i \mid x_i \in X \wedge x_i \leq x_{i+1},\ i \in \{1, \ldots, n\}\}$ and $Y' = \left\{z_i \mid z_i = y_{index\_of(x_i)} \wedge y_i \in Y,\ i \in \{1, \ldots, n\}\right\}$. Based on such two lists, we define the following agreement function[4]

$$\text{pos\_agr}(A, B, i, j) = \begin{cases} 1 & \text{if } x_i \neq x_j \wedge y_i < y_j \wedge x \in A \wedge y \in B \\ 0 & \text{otherwise,} \end{cases}$$

that tells us whether the ordering of two documents is consistent across the two judging sets. Thanks to this we can now define the agreement score as the ratio of consistent document pairs over all possible pairs:

$$\left(\sum_{i=1}^{n} \sum_{j=i+1}^{n} \text{pos\_agr}\left(X', Y', i, j\right)\right) \cdot \binom{n}{2}^{-1}.$$

Note that this is not a symmetric measure, but it rather computes agreement of $Y$ ratings as compared to $X$ considered the baseline judgments. This measure computes the number of agreement pairs between the two datasets. That is, if a document $w$ has a higher relevance judgment score than a document $z$ according to judgments in $X$, we would like the same order $w \geq z$ to be maintained in $Y$. That is, the relevance judgment score of $w$ should be higher than $z$ according to $Y$ judgments.

*4.2.2 Comparison with other scales.* Figure 5 shows the complementary cumulative distribution function (showing how often

[4]We consider $y_i < y_j$ rather than $x_i \leq x_j$ as we assume $X$ to use a coarser-grained scale (e.g., binary) as compared to $Y$ (e.g., S100).

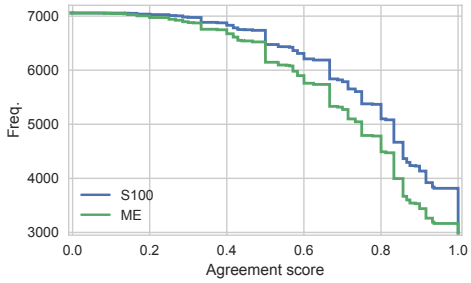


**Figure 5: Complementary cumulative distribution function of pairwise agreement as defined in Section 4.2.1 of S100 and ME with TREC.**

agreement is above a given value) of pairwise agreement for S100 and ME with respect to TREC binary judgments. The ME series is another representation of the data in [15, Figure 6]. The comparison highlights that agreement levels in S100 are higher than ME.

Figure 6 shows the topics ordered by another standard measure of agreement, Krippendorff's $\alpha$ [13]. We can make the following observations:

- Agreement scores for judgments collected with S100 are substantially higher than those collected with ME.
- There is some consistency across S100 and ME in the sense that topics with high/low agreement tend to be the same.
- Agreement over TREC non-relevant documents is higher as compared to relevant ones.
- Agreement on the non-relevant documents as compared to agreement on all the documents is similar in the two figures, whereas agreement on the relevant documents as compared to agreement on all the documents is higher in S100 than in ME (the green "TREC: 1" series is "pulled up" in the S100 chart). In other terms, S100 improves, w.r.t. ME, $\alpha$ agreement on the relevant documents.

## 4.3 IR System Ranking Correlation

Finally, we computed Kendall $\tau$ correlation of IR systems ranked by effectiveness computed using judgments collected over different relevance scales. Figure 7 shows the IR system ranking correlation using NDCG@10 [11] when using binary judgments as compared to S100 (a), using binary judgments as compared to ME (b); and using S100 as compared to ME (c). Each dot is a system and the charts show its NDCG@10 values over two different scales. We can observe that, while all judgments result in high system ranking correlation values, the best correlation is obtained when comparing S100 and ME. This demonstrates how S100 lead to results similar to ME by providing assessors the flexibility to judge document relevance on a fine-grained basis. This is also explained by the fact that S100 and ME have been collected following the same crowdsourcing setup while the TREC and S4 did not use crowdsourcing. Looking at how S100 and ME compare with TREC (Figure 7 a and b) we can see that while correlation values are similar, NDCG@10 scores obtained using S100 are more consistent with those obtained with TREC labels whereas ME judgments tend to result in lower NDCG scores.

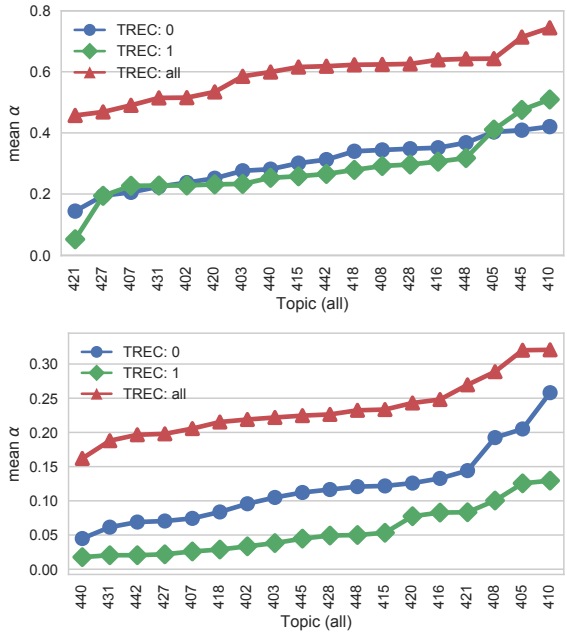


**Figure 6: Agreement ($\alpha$) of the individual topics, in S100 (above) and ME (below).**

# 5 ROBUSTNESS TO FEWER JUDGMENTS

In this section we study how different relevance scales behave in terms of robustness to fewer judgments. That is, we look at how crowdsourced relevance label quality decreases as compared to editorial judgments by experts like TREC and Sormunen's S4 assessors. In detail, we study two kinds of robustness:

- Shorter HITs: including fewer documents to be judged in a HIT so that each worker has the option to do less if they wish to.
- Fewer assignments per document: using fewer workers judging the same document, and averaging their judgments.

We measure robustness by observing how pairwise agreement with TREC and S4 decreases.

## 5.1 Fewer Documents per HIT

In the crowdsourcing setup used to create the S100 and ME collections each worker is required to judge 8 documents in one HIT. When using fewer documents per HIT, we assume we could lose on training effects (i.e., workers becoming proficient in the judging task) with the benefit of work flexibility.

Figure 8 shows how pairwise agreement varies when using shorter HITs (i.e., looking at judgment quality based on the document position in the HIT). For any HIT length, the pairwise agreement of individual judgments is higher for S100 than ME with an increasing gain in agreement the longer the HIT.

## 5.2 Fewer Judgments per Document

In both S100 and ME, 10 judgments per document have been collected. When using fewer assignments, as it is expected, the quality of the aggregated judgments decreases. We analyze this by showing

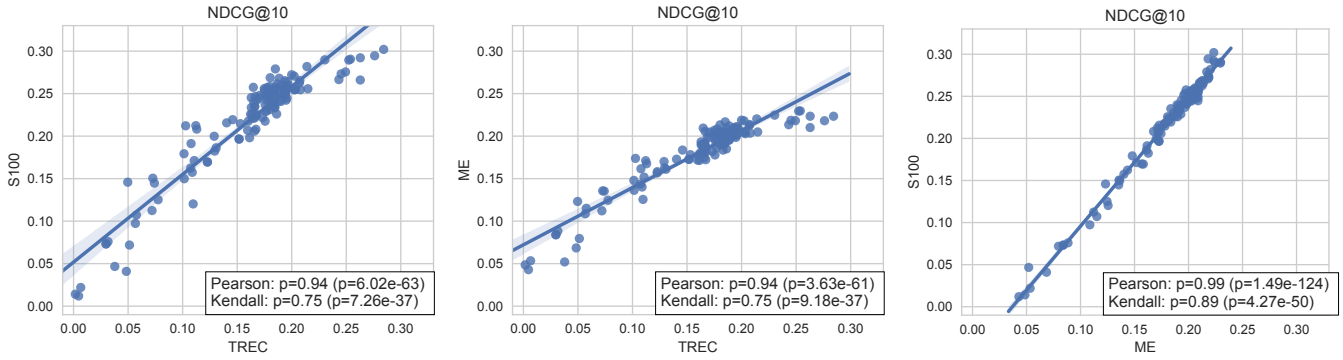


**Figure 7: NDCG@10 scores for TREC-8 runs and judgments collected over different scales: TREC, ME and S100.**

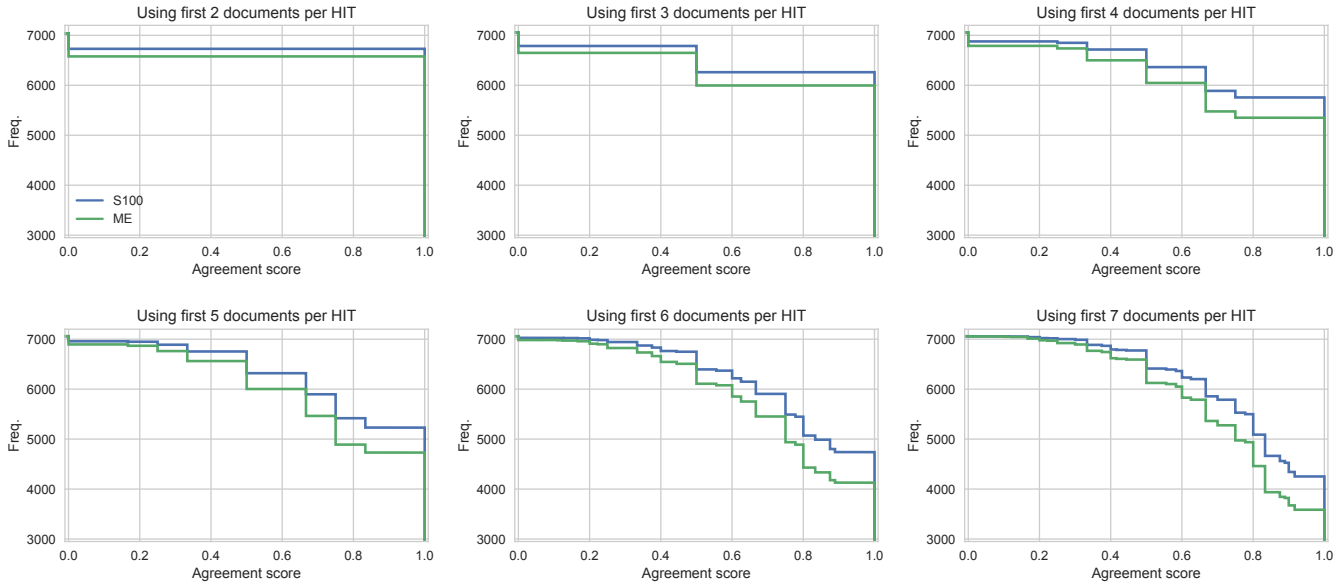


**Figure 8: Complementary cumulative distribution function of pairwise agreement as defined in Section 4.2.1 of S100 and ME with TREC, when using the first $i \in \{2, \ldots, 7\}$ documents for each HIT ($i = 8$ shown in Figure 5).**

how aggregated pairwise agreement decreases when using a random subsample of the 10 judgments in Figure 9: the figure reports the pairwise agreement average values, for each topic, over 100 random repetitions. For all sizes of the subsample, S100 median pairwise agreement is higher for S100 than ME.[5] These results show a higher robustness to fewer assignments of S100 as compared to ME, making it a more economically viable scale to use to collect crowdsourced relevance judgments.

## 6 RUNNING OUT OF VALUES

As recalled above, the ME scale has the advantage that the assessor never "runs out of values", neither (i) at the scale extremes (which are unbounded) nor (ii) inside the scale (which is continuous). This advantage is lost when using limited scales as S100, and it is of course further compounded for scales with a lower amount of values like S4 and TREC. In this section we aim at understanding if these potential problems have actually been a practical constraint for the judges using the S100 scale. An initial analysis can be performed comparing the number of "back" actions crowd workers performed which indicate a desire to change or look at their previously judged documents. This value is much higher in S100 than in ME. In ME the number of single "back" actions was 106, and the number of two or more "back" actions was 9 [15, Table I], whereas for S100 these two figures are 113 and 182 respectively: although the numbers are still very limited (more than 95% of S100 workers did not use the "back" button at all), the differences are noticeable and might be ascribed to a higher difficulty in finding the "right" relevance score.

[5] We observed the same result when computing pairwise agreement against Sormunen's S4 judgments but we did not include the figure for space limitations.

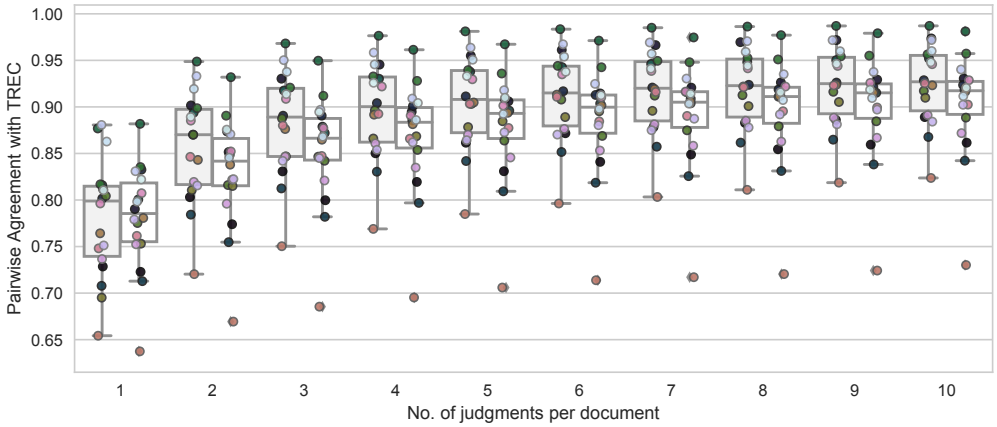


**Figure 9: Pairwise agreement of S100 (boxplot on the left) and ME (boxplot on the right) with TREC, when sampling randomly (mean over 100 repetitions) $i \in \{1, \ldots, 10\}$ judgments for each document. Each dot is a topic.**

**Table 1: The number of cases with exactly $k$ 0s or 100s in the same HIT, and the corresponding number of potential run-out-of-values cases in the S100 dataset.**

| $k$ | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Tot |
|---|---|---|---|---|---|---|---|---|---|
| 0s | *2355* | *803* | 694 | 689 | 736 | 786 | 641 | 355 | |
| 100s | *3796* | *1808* | 846 | 398 | 133 | 55 | 19 | 4 | |
| $k-1$ | | | 1 | 2 | 3 | 4 | 5 | 6 | |
| 0s* | 0 | 0 | 694 | 1378 | 2208 | 3144 | 3205 | 2130 | 12759 |
| 100s* | 0 | 0 | 846 | 796 | 399 | 220 | 95 | 24 | 2380 |

In the next two subsections we present a more detailed analysis, addressing the two issues (i) and (ii).

## 6.1 Reaching the Scale Boundaries

Table 1 shows in the first part the number of HITs with "boundary judgments", i.e., with exactly $k$ 0s or 100s in the S100 dataset. The HITs without or with only one boundary judgment (i.e., only one 0 or one 100, in italics in the table) do not create any potential problem; instead, the boundary judgments after another boundary judgment (i.e., in the same HIT, one or more 0s after a first 0, or one or more 100s after a first 100) might be cases in which the worker could have used a lower or higher value if available. So, the HITs with at least two boundary values (the following columns) are those in which, at least potentially, the worker "ran out of values" at each extreme of the scale. A lower (higher) value, if available, could have been selected for each of the $k-1$ "boundary judgments" after the first one. The numbers of such "boundary judgments following other boundary judgment(s)" are quantified in the lower part of the table: these are obtained multiplying by $k-1$ the figures in the previous two rows (for example, the 2'208 value is obtained as 736x(4-1): in 736 HITs the workers used 0 for 4 times, and the last three in each HIT are candidates for "run-out-of-value" cases).

To provide an understanding of the frequency of the problem, let us remember that we had a total of 7'059 HITs, each one containing 8 documents. Since the first expressed judgment in each unit can not be preceded by another (boundary) judgment, we have 7'059x7=49'413 judgments that could have manifested the problem. Of those, the problem manifests for a total of 12'759+2'380=15'139 cases (31%). Of course this is not negligible: in almost one case out of three a worker might have been restricted in expressing the true intended judgment. However, this also means that in 69% of the expressed judgment we can say that the worker was not restricted by the boundaries of the S100 scale. Moreover, these 31% of cases are only potential problems, as it might well be that the worker intended to express exactly the same judgment and did not actually run out of values. Therefore, we further analyzed these potential run-out-of-values cases in two ways.

First, we looked in our S100 dataset what fraction of the boundary judgments 0 (or 100) expressed by a worker in a HIT after the first boundary judgment corresponds to a document that has a strictly lower (higher) aggregated score. These are cases in which the worker ran out of values, assuming that the intended score corresponds to the aggregated one. This happened 7'523 cases, namely 50% of the 15'139 potential problematic cases, or 15% of the total 49'413 judgments expressed.

Second, we looked in the ME dataset how many of the 15'139 potential run-out-of-value cases received an ME score that was lower (for the 0s), or, respectively, higher (for the 100s) than the first corresponding boundary judgment in the unit. These are cases in which the worker ran out of values assuming that the judgment expressed by the ME worker in the corresponding unit was exact. This happened for 4'309 cases, namely 28% of the 15'139 potential problematic cases or 9% of the total 49'413 judgments expressed.

So, the two analyses roughly agree that in only around one out of ten cases the bounded scale seems to have indeed limited the assessor, and therefore in about 90% of the expressed judgments the S100 scale did not create any obstacle to judgment expression.

## 6.2 Discrete vs. Continuous Scale

The second situation in which the S100 scale could constrict judgment expression is when contiguous values are selected, thus making impossible for the worker to select another, different, value in between in the following judgments in the same HIT. We counted

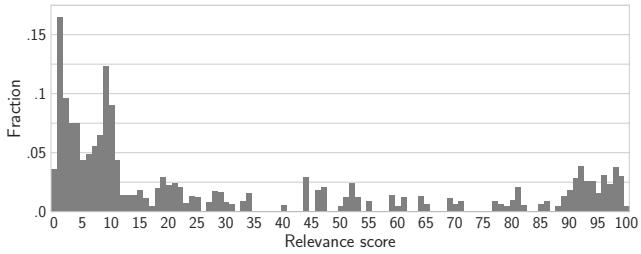


**Figure 10: Fraction of expressed judgments, for each value of the S100 scale, that might have been constrained by previously expressed judgments in the same HIT.**

in our S100 dataset how many scores $x$ were preceded by both $x$ and $x+1$ (or by both $x-1$ and $x$) in the same HIT. These are the cases in which, potentially, the worker could not give a $y \in ]x-1, x[$ (or $y \in ]x, x+1[$) value because of the discrete scale. There were 1'911 such cases, out of the 7'059x6=42'354 possible ones (as we need to count starting from the third judgment in each HIT), which is less than 5%. Notice again that this fraction is consistent with the number of 'back' actions. Moreover, the vast majority of these cases (around 1'500) concern judgments between 0 and 10, which could be considered not critical (as it is probably more important to focus on the "relevant" end of the scale rather than on the "not relevant" end). This is confirmed by Figure 10 that shows, for each value between 0 and 100, the fraction of judgments at a given level that may have been affected by previous judgments given at the same level. These values indicate the percentage of cases in which there may be a limitation because of the use of S100 as compared to ME which would allow to give a slightly higher or lower judgment score as compared to previous ones. Note that this is an upper bound of such expressiveness limitation as assessors may have assigned the same score multiple times on purpose. From Figure 10 we can also observe that most problematic cases are, as expected, at the boundaries of the scale but also that such cases are not prevalent (about 5% of judgments are affected). More potentially constrained judgments are present at the lower end of the S100 scale. This should be less problematic than constrains at the upper end of the scale as we can expect more score ties for not relevant documents.

In summary, taking into account that these are only potential problems, since it is possible that the worker indeed intended to express the very same score again, we can conclude that these do not seem worrying problems in practice and that the theoretical constraints imposed by the S100 scale on judgments expression did not significantly obstacle the workers in practice.

## 7 JUDGMENT TIME

We compared the time required by crowd workers to assess documents using S100 and ME considering that the experimental design was consistent across the two studies with the only difference being the way relevance was expressed by assessors (i.e., a number versus a slider from 0 to 100).

The dotted series in Figure 11 show the mean time taken by crowd workers to assess documents, based on the order in which they were presented. The cumulated time shows that S100 leads to slightly quicker judgments than ME for the first 5 documents.

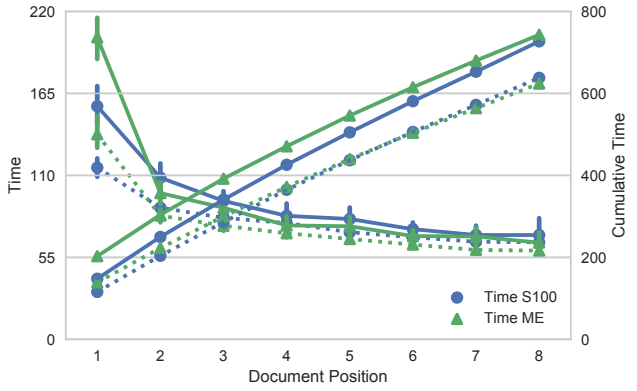


**Figure 11: Mean time (seconds) for expressing relevance judgments, over the 8 positions in each HIT, considering either only the first HIT for each worker (straight lines) or considering all HITs (dotted lines). All times difference are significant (Wilcoxon signed-rank test $p < 0.01$).**

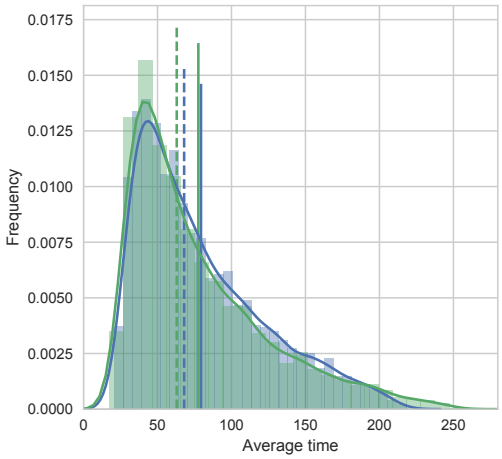


**Figure 12: Judging time distributions, combined for all topics, for S100 and ME. Average (vertical lines) and median (dashed vertical lines) time is also shown.**

Starting from the 6th document in the HIT, the use of the ME scale leads to overall faster judgments. This can be explained by the fact that workers are probably at first disoriented by the uncommon ME scale, but they become more efficient in using it as they progress completing more judgments. The time behavior is quite stable across workers and topics (as it can be seen by the small quartile bars in Figure 11). Moreover, time differences are quite small, though always statistically significant, for each document position in the HIT: apart from the first position, the differences are around 5s for a total judgment time of 50-100s.

Figure 12 shows in more detail the distributions of judging time for both S100 and ME: the two are rather similar, with ME having a slightly longer tail. The behavior does not significantly change across individual topics (not shown for brevity). Given the small and constant time difference, a possible explanation is the presence of

the slider, which might require a small longer amount of time than inserting a number in a text box, especially on mobile devices as shown by Gadiraju et al. [6]. Further analysis is needed to confirm this conjecture, and we leave that to future work.

To further study the learning effects, we repeated the same analysis considering only the first HIT performed by each worker (remember that workers could redo the task, on a different topic). Going back to Figure 11, the straight, not dotted, lines represent only these first HITs, for both S100 and ME. We notice three main variations with respect to the dotted lines: (i) overall, average times are higher, indicating that indeed some learning effect is present and workers become more efficient after the first HIT; (ii) the time difference on the first expressed judgment is larger, thus confirming that starting with the ME scale is somehow more difficult than with the S100 one; and (iii) on this data, the ME cumulative curve stays above the S100 one, meaning that the disadvantage cumulated on the first document by ME can not be recovered even after 8 judgments are judged.

Overall, we can conclude that time does not seem a critical factor when choosing between S100 and ME: differences are small and a longer time due to learning effects on ME tends to be compensated after some documents are judged.

## 8 CONCLUSIONS

In this paper we presented a systematic study comparing the effects of different relevance scales on IR evaluation. We have shown many advantages of the S100 scale as compared to coarse-grained scales like binary and S4 and to unbounded scales like ME. S100 preserves many of the advantages of ME like, for example, allowing to gather relevance judgments that are much more fine-grained than the usual binary or 4-value scales. Assessors use the full spectrum, although sometime with a preference for scores that are a multiple of ten. S100 has also demonstrated advantages over ME in terms of agreement with judgments collected on a binary and four level scales. This can be explained by the fact that ME requires a step of score normalization which makes judgments less comparable across assessors and topics. On the other hand, S100 leads to more similar judgments (i.e., higher agreement) to the classic binary and four level scales. S100 has shown to be more robust than ME in terms of less assessors per documents (to be aggregated, as typically done for crowdsourced relevance judgments) and to less documents per assessor thus giving the freedom to crowd workers to perform few or many judging tasks. Our results show that the potential constraints in judgment expression that the S100 scale might create with respect to the complete freedom of ME almost do not occur in practice since about 90% of the judgments did not suffer from this problem. The S100 scale also seems easy to learn for the workers and turns out to be faster than ME for short HITs with 5 or less documents to be judged, and of comparable speed for longer HITs. Overall, our results show that S100 is an effective, robust, and usable scale to gather fine-grained relevance labels.

**Acknowledgements.** This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 732328.

## REFERENCES

[1] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Inf. Process. Manage.* 48, 6 (2012), 1053–1066. https://doi.org/10.1016/j.ipm.2012.01.004

[2] Olivier Chapelle, Donald Metlzer, Ya Zhang, and Pierre Grinspan. 2009. Expected Reciprocal Rank for Graded Relevance. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM '09)*. ACM, New York, NY, USA, 621–630. https://doi.org/10.1145/1645953.1646033

[3] Charles LA Clarke, Nick Craswell, and Ian Soboroff. 2004. Overview of the TREC 2004 Terabyte Track.. In *TREC*, Vol. 4. 74.

[4] Kevyn Collins-Thompson, Craig Macdonald, Paul Bennett, Fernando Diaz, and Ellen M Voorhees. 2015. *TREC 2014 web track overview.* Technical Report. MICHIGAN UNIV ANN ARBOR.

[5] Carsten Eickhoff. 2018. Cognitive Biases in Crowdsourcing. In *WSDM 2018*. To appear.

[6] Ujwal Gadiraju, Alessandro Checco, Neha Gupta, and Gianluca Demartini. 2017. Modus Operandi of Crowd Workers: The Invisible Role of Microtask Work Environments. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 3, Article 49 (Sept. 2017), 29 pages. https://doi.org/10.1145/3130914

[7] George A Gescheider. 2013. *Psychophysics: the fundamentals*. Psychology Press.

[8] Mehdi Hosseini, Ingemar J. Cox, Natasa Milic-Frayling, Gabriella Kazai, and Vishwa Vinay. 2012. On Aggregating Labels from Multiple Crowd Workers to Infer Relevance of Documents. In *Advances in Information Retrieval - 34th European Conference on IR Research, ECIR 2012, Barcelona, Spain, April 1-5, 2012. Proceedings*. 182–194. https://doi.org/10.1007/978-3-642-28997-2_16

[9] Quan Huynh-Thu, Marie-Neige Garcia, Filippo Speranza, Philip Corriveau, and Alexander Raake. 2011. Study of rating scales for subjective quality assessment of high-definition video. *IEEE Transactions on Broadcasting* 57, 1 (2011), 1–14.

[10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (Oct. 2002), 422–446. https://doi.org/10.1145/582415.582418

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

[12] Jiepu Jiang, Daqing He, and James Allan. 2017. Comparing In Situ and Multidimensional Relevance Judgments. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '17)*. ACM, New York, NY, USA, 405–414. https://doi.org/10.1145/3077136.3080840

[13] Klaus Krippendorff. 2007. Computing Krippendorff's alpha reliability. *Departmental papers (ASC)* (2007), 43.

[14] Eddy Maddalena, Marco Basaldella, Dario De Nart, Dante Degl'Innocenti, Stefano Mizzaro, and Gianluca Demartini. 2016. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[15] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On Crowdsourcing Relevance Magnitudes for Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 35, 3 (2017), 19.

[16] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Fourth AAAI Conference on Human Computation and Crowdsourcing*.

[17] Tetsuya Sakai. 2007. On the Reliability of Information Retrieval Metrics Based on Graded Relevance. *Inf. Process. Manage.* 43, 2 (March 2007), 531–548. https://doi.org/10.1016/j.ipm.2006.07.020

[18] 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 American Society for Information Science and Technology* 58, 13 (2007), 1915–1933. https://doi.org/10.1002/asi.20682

[19] Eero Sormunen. 2002. Liberal Relevance Criteria of TREC -: Counting on Negligible Documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '02)*. ACM, New York, NY, USA, 324–330. https://doi.org/10.1145/564376.564433

[20] Rong Tang, William M Shaw Jr, and Jack L Vevea. 1999. Towards the identification of the optimal number of relevance categories. *Journal of the Association for Information Science and Technology* 50, 3 (1999), 254.

[21] 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 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '15)*. ACM, New York, NY, USA, 565–574. https://doi.org/10.1145/2766462.2767760

[22] Matteo Venanzi, John Guiver, Gabriella Kazai, Pushmeet Kohli, and Milad Shokouhi. 2014. Community-based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web (WWW '14)*. ACM, 155–164. https://doi.org/10.1145/2566486.2567989