How Many Crowd Workers Do I Need? On Statistical Power when Crowdsourcing Relevance Judgments

KEVIN ROITERO, DAVID LA BARBERA, and MICHAEL SOPRANO, University of Udine, Italy
GIANLUCA DEMARTINI, The University of Queensland, Australia
STEFANO MIZZARO, University of Udine, Italy
TETSUYA SAKAI, Waseda University, Japan

To scale the size of Information Retrieval collections, crowdsourcing has become a common way to collect relevance judgments at scale. Crowdsourcing experiments usually employ 100–10,000 workers, but such a number is often decided in a heuristic way. The downside is that the resulting dataset does not have any guarantee of meeting predefined statistical requirements as, for example, have enough statistical power to be able to distinguish in a statistically significant way between the relevance of two documents.

We propose a methodology adapted from literature on sound topic set size design, based on t-test and ANOVA, which aims at guaranteeing the resulting dataset to meet a predefined set of statistical requirements. We validate our approach on several public datasets.

Our results show that we can reliably estimate the recommended number of workers needed to achieve statistical power, and that such estimation is dependent on the topic, while the effect of the relevance scale is limited. Furthermore, we found that such estimation is dependent on worker features such as agreement. Finally, we describe a set of practical estimation strategies that can be used to estimate the worker set size, and we also provide results on the estimation of document set sizes.

CCS Concepts: • **Information systems** → **Evaluation of retrieval results;**

Additional Key Words and Phrases: Crowdsourcing, statistical analysis, relevance judgments

ACM Reference format:
Kevin Roitero, David La Barbera, Michael Soprano, Gianluca Demartini, Stefano Mizzaro, and Tetsuya Sakai. 2023. How Many Crowd Workers Do I Need? On Statistical Power when Crowdsourcing Relevance Judgments. *ACM Trans. Inf. Syst.* 42, 1, Article 21 (August 2023), 26 pages.
https://doi.org/10.1145/3597201

1 INTRODUCTION

Test collections are a mechanism that can be used to reliably measure the effectiveness of **Information Retrieval (IR)** systems. When building a new test collection, the most expensive step (in terms of both human time and money) is the process of collecting relevance judgments for each

This work is partially supported by the Australian Research Council (ARC) Training Centre for Information Resilience (Grant No. IC200100022).

Authors' addresses: K. Roitero, D. La Barbera, M. Soprano, and S. Mizzaro, University of Udine, Udine, Italy; emails: kevin.roitero@uniud.it, david.labarbera@uniud.it, michael.soprano@uniud.it, mizzaro@uniud.it; G. Demartini, The University of Queensland, Brisbane, Australia; email: demartini@acm.it; T. Sakai, Waseda University, Tokyo, Japan; email: tetsuya@waseda.jp.

1046-8188/2023/08-ART21 $15.00
https://doi.org/10.1145/3597201

(topic, document) pair belonging to the pooled set of retrieved documents. Since this approach does not scale up, researchers proposed to use crowd workers (e.g., from Amazon Mechanical Turk) as a valid alternative to classical human assessors and discussed the effectiveness of such an approach [1, 31, 33, 34, 44, 46, 54, 62].

The design of the crowdsourcing tasks aimed at collecting relevance judgments is left to researchers and practitioners, who select and design the assessment task with a focus on the number of documents to be judged based on pooling strategies; conversely, the number of workers n assigned to assess each document is often set in a heuristic way using a rule of thumb. Finally, given the associated costs, such a number n of workers is usually minimized. This approach, when compared to using classical "perfect" relevance assessors, has two consequences. A first one is that, as crowdsourcing experts know well, if n is too low, then the assessments could be inaccurate. This is a crucial aspect, since misclassification of document relevance may lead to incorrect values for the effectiveness metric, which would result in an inaccurate evaluation of the systems' effectiveness. However, there is a second consequence, which is the focus of this article: The annotated datasets generated by those crowdsourcing tasks cannot guarantee to fulfill a predefined set of statistical requirements. For instance, the dataset may not possess adequate *statistical power* to differentiate each pair of documents in a *statistically significant* way. In this article, we propose a methodology to ensure that a group of items that is judged by multiple workers can be distinguished in a statistically significant way. This is another crucial aspect to recognize, because it enables us to identify any genuine differences in document relevance.

Our work is based on the assumption that, at least in some contexts, it is important to distinguish with statistical significance between the relevance of two or more documents. Although this assumption might not hold in every situation, this article is intended to be read with such an assumption in mind, as our research focuses on scenarios where differentiating between documents with statistical significance is of primary importance. One such example where our assumption holds is when collecting preference judgments [7–10, 61], where it is necessary to distinguish between items that have similar levels of relevance. Another case is metrics-free evaluation [14], where effectiveness is calculated by directly computing a preference between ranked lists. These problems are notable, because they have many practical applications, also beyond test collections; for instance, social media platforms face the challenge of distinguishing items reliably (i.e., with statistical significance).

There are some differences and connections between the approach proposed in this article and the typical approach employed in test collections, as well as in Sakai's research [48–50]. In traditional offline ad hoc retrieval evaluation, the effectiveness of two (or more) systems is compared by examining whether their mean effectiveness scores differ, with the topic set of a given test collection being viewed as a sample from the underlying population of all possible topics. The number of topics required to achieve a specified statistical power is determined by the topic set size design principle [48–50]. In this article, we present a novel extension of this principle by taking into account not only the system and topic sets but also the underlying set of judgments that are used to derive the effectiveness scores for each system–topic pair. This approach aims to create a *statistically reliable* test collection, where we employ a population of users who rate a set of documents, and our crowd workers are treated as being a sample from the underlying user population. We aim to hire a sufficient number of crowd workers so we can detect a substantial true difference in relevance between two (or more) documents with a designated statistical power, which allows us to accurately measure systems' effectiveness. Thus, we seek to address the question of how many assessor labels should be collected per document without relying on heuristics.

To summarize, whereas the number of topics is a concern for IR researchers and practitioners comparing systems using an existing test collection, our work is intended for test collection builders seeking to determine the number of document relevance labels to collect to ensure high statistical power when two or more documents have truly different degrees of relevance.

Thus, in this work, we propose a methodology to estimate the number of crowd workers recommended to annotate a test collection that achieves a given statistical power.[1] Our method extends prior work based on the t-test and one-way ANOVA, and it allows the researchers and practitioners to estimate beforehand the number of workers n to employ in a crowdsourcing task to obtain as a result an annotated dataset with a minimum statistical power that is guaranteed by design.

In particular, we build on top of the works done by Sakai [48–50], who studied the number of topics recommended to achieve statistical power in the setting of offline evaluation of retrieval systems using test collections. While we share some similarity with Sakai's work, our approach has important differences. Sakai's work considered a system-by-topic complete matrix and computed, for a given evaluation metric (such as precision, recall, etc.), the number of topics recommended to distinguish in a statistically significant way two or more retrieval systems. In this work, instead, we adapt the matrix detailed above focusing on the context of crowdsourcing relevance judgments. We build document-by-worker and worker-by-document sparse matrices for a given assessment scale (such as binary scale, interval scale, etc.). The first matrix (document-by-worker) is used to compute the number of workers recommended to distinguish two or more documents for a considered relevance scale, while the second (worker-by-document) is used to compute the number of documents recommended to distinguish in a statistically significant way two or more crowd workers. Both approaches have important practical applications; while the former approach is useful to understand if two documents that appear to have received two distinct aggregated judgments are actually different in terms of relevance or not, the latter is useful to understand whether two workers are different in terms of quality, distribution of judgments, and so on. The former case might be the most interesting one for the IR community, as the final aim when crowdsourcing relevance judgments is indeed on the documents scores. We therefore primarily focus on the document-by-worker setting in this article.

We experimentally evaluate our proposed approach in the setting of crowdsourcing relevance judgments, using multiple publicly available datasets. Our results show that the proposed methodology can provide a reliable estimation of the number of workers recommended to distinguish two documents in a statistically significant way, and of the number of documents recommended to distinguish two workers in a statistically significant way. Furthermore, we provide researchers and practitioners with a methodology to reliably compute such an estimation before deploying a complete crowdsourcing experiment. All the data and code used in this article is made publicly available and can be downloaded at https://github.com/KevinRoitero/how-many-crowd-workers.

2 BACKGROUND AND RELATED WORK

We summarize the literature about statistical power in test collections design and in crowdsourcing experimentation.

2.1 Statistical Power of Test Collections

Statistical power of test collections has received some attention by the IR community. Nelson [36] studied statistical tests with a focus on effect sizes and real-life impact of the design of test collections. Smucker et al. [56] compared five commonly used tests, namely, the Student's paired t-test, the Wilcoxon signed rank test, the sign test, the bootstrap, and the Fisher's randomization test. They study the practical differences between those tests when observing differences between runs

[1] We use the term "recommended" to be aligned with Sakai's work; however, the term should be interpreted as "required."

in the TREC 3, 5, 7, and 8 test collections. Webber et al. [60] considered different techniques to estimate the system population variance recommended to estimate the number of topics needed in a test collection to achieve statistical power and proposed a hybrid methodology used to demonstrate that in test collections it is generally better, in terms of statistical power, to evaluate a large number of topics with a shallower pool rather than a small number of topics with a deeper pool. More recently, Ferro and Sanderson [20] investigated the significance of statistical tests used to compare and measure difference in the effectiveness of retrieval systems. Sakai [48–50] focused on the effect size needed to estimate the number of topics needed in a test collection and provided a sound and reliable methodology to estimate the number of topics recommended in a test collection, depending on the experimental settings and hypotheses to be tested [50]. Related to this, other works investigated the number of topics recommended in test collections: Carterette et al. [5] investigated the minimal dimension of a test collection required to reliably estimate IR system effectiveness differences. Buckley and Voorhees [4] focused on the differences in evaluation measures when different topic set sizes are used, while Berto et al. [3], Guiver et al. [24], Roitero et al. [43], and Roitero et al. [47] investigated the impact of using a different set of topics to compute system effectiveness. Other work focused on statistical power in relation to judgment pool depth [55]: Sakai and Kando [51] and Sakai and Mitamura [52] performed experiments varying the pool depth, Zobel [64] and Cormack and Lynam [12] measured the reliability of experimental results when considering different pooling strategies and pool depths. Some recent work investigated the tests used to compute statistical significance for information retrieval: Parapar et al. [40] characterized the behavior of significance tests in retrieval ranking tasks, while Urbano et al. [59] focused on the system dependency when performing stochastic simulation of evaluation data used to compare statistical significance tests.

Compared to previous work, we look instead at how statistical methods can inform the choice of how many distinct workers are needed to judge the relevance of a document in a way it can be distinguished from others. Note that to validate our approach, we rely on literature work that employed a crowd of workers recruited with the aim of assessing the relevance of a set of documents. In the following, using crowdsourcing terminology, we refer to the crowd of workers recruited using the word “workers,” which in this setting is a synonym for “assessor.”

2.2 Statistical Power of Crowdsourcing Experiments

As opposed to what happened to other disciplines such as IR, a formal study of statistical power and significance in the setting of crowdsourcing experiments received little attention [17]. While a formal study detailing the effect and consequences of an experimental design is missing, some work considered in a central way statistical power and significance for crowdsourcing experiments.

Kittur et al. [29] pointed out, in their influential work, the relationship between a good experimental formulation and obtaining good results from the crowd workers. Ribeiro et al. [41] proposed a tool to conduct Mean Opinion Score tests to evaluate signal processing methods using crowdsourcing and considered statistical significance of the crowd sample employed. Behrend et al. [2] compared, considering statistical significance, the viability of using crowdsourcing platforms to recruit participants as opposed to university students when conducting surveys for behavioral research. Eickhoff and De Vries [17] considered statistical significance to increase the robustness of crowdsourcing tasks by identifying malicious workers. Landy et al. [32] compared the research outcome of 15 research groups on a common subject and studied how the design choices influenced the significance of the results.

Compared to previous work, we look instead at how statistical methods can be used to decide how many documents should be assessed by workers to be able to distinguish them.

3 AIMS AND RESEARCH QUESTIONS

In this work, we follow an approach different from those of previous works in considering statistical power for crowdsourcing. In more detail, we provide a general methodology that can estimate the number of workers (documents) recommended to have enough statistical power to be able to distinguish two documents (workers) when collecting relevance judgments using crowdsourcing. Furthermore, we experimentally validate our approach using real data collected when crowdsourcing relevance assessments, focusing on the following research questions:

RQ1 Can we reliably estimate the recommended number of workers needed to achieve statistical power for distinguishing between two documents when crowdsourcing relevance judgments? Is such estimation topic-dependent?
RQ2 Is the estimation of the recommended number of workers needed to achieve statistical power dependent on the relevance of the documents?
RQ3 How does the estimation of the recommended number of workers change when different experimental settings are employed? In more detail, how do workers' features such as arrival time or quality impact such an estimation?
RQ4 Which is the theoretical highest and lowest estimation of the recommended number of workers needed for a test collection to satisfy a set of statistical constraints?
RQ5 Can we identify an effective strategy that can be applied in practice to estimate beforehand the recommended number of workers needed to achieve statistical power when crowdsourcing relevance judgments?
RQ6 Can we reliably estimate the recommended number of documents needed to achieve statistical power for distinguishing between two workers when crowdsourcing relevance judgments?

The results of our work contribute to the body of knowledge that aims to design a more robust, sound, and engineered approach to the building of a test collection; furthermore, to the best of our knowledge, our work is the first one addressing statistical power in the design of crowdsourcing experiments to collect relevance judgments.

4 METHODOLOGY

To make this work self-contained, we first introduce the mathematical concepts and the methodology used to perform the experiments detailed in this article. We refer to the works by Sakai [48, 49, 50] for additional details.

Suppose we have n workers and $m \geq 2$ documents to be compared; we use the one-way ANOVA model as follows: Let x_{ij} be the assessment score for the ith document given by the jth worker according to some relevance scale; let us assume that $\{x_{ij}\}$ are independent and $x_{ij} \sim \mathcal{N}(\mu, \sigma^2)$, i.e., the x_{ij} scores follow a normal distribution with mean μ and, according to the homoscedasticity assumption, a common variance σ^2. Before moving on with the model, let us make a remark on the normality assumption. Previous work has shown that crowdsourced relevance judgments are not normally distributed [44, 46]. Nevertheless, while the mathematical derivation of the t-test-based power analysis starts from the normality assumption, it is known that t-test is quite robust to violations of this assumption in practice [37]. Especially when the sample size is sufficiently large (say, 30), sample means approximately obey normal distribution regardless of the distributions of the individual scores (due to the Central Limit Theorem). Thus, the robustness of the t-test can be demonstrated by comparing t-test p-values with randomization-test p-values, where the latter does not rely on the normality assumption. In this setting, it is known that the p-values are generally similar (though not identical). Hence, our approach is expected to be reasonably robust to normality assumption violations [50].

We then model our scenario according to ANOVA without replication model, defined as $x_{ij} \sim \mu + a_i + \epsilon_{ij}, x_{ij} \sim \mathcal{N}(0, \sigma^2)$, where μ is the population mean, σ^2 is the common variance, a_i is the ith document effect, and ϵ_{ij}, x_{ij} is the error term that obeys $\mathcal{N}(0, \sigma^2)$. In other words, we are assuming that the model effects are both additive and linearly related to x_{ij}.[2]

We can then model the m-by-n document-by-worker matrix as follows: We denote with: $D = \{d_1, \ldots, d_m\}$ the set of documents, $W = \{w_1, \ldots, w_n\}$ the set of workers, x_{ij} the assessment of the jth worker for the ith document if present, $\perp$ (or a placeholder, otherwise), $W_j = \{w_j \in W | x_{ij} \neq \perp\}$ the set of workers that assessed the ith document, $D_i = \{d_i \in D | x_{ij} \neq \perp\}$ the set of documents that have been assessed by the jth worker, $\overline{x}_{i\bullet}$ the sample mean for the ith document. Then, we can represent the m-by-n matrix as

$$\begin{array}{c|ccccc|c} & w_1 & w_2 & w_3 & \dots & w_n & \\ d_1 & x_{11} & x_{12} & x_{13} & \dots & x_{1n} & \overline{x}_{1\bullet} \\ d_2 & x_{21} & x_{22} & x_{23} & \dots & x_{2n} & \overline{x}_{2\bullet} \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ d_m & x_{m1} & x_{m2} & x_{m3} & \dots & x_{mn} & \overline{x}_{m\bullet} \end{array}$$

We can then compute the residual V_{E1} (only for $x_{ij} \neq \perp$) as:

$$V_{E1} = \frac{\sum_{i=1}^{m} \sum_{j=1}^{n} (x_{ij} - \overline{x}_{i\bullet})^2}{\sum_{i=1}^{m} |D_i| - m}.$$

This is an unbiased estimate of the common variance σ^2; the lower the variance (thus the disagreement among workers on each document), the lower the number of workers needed to achieve statistical power. Then, we can compute the number of recommended workers by plugging the residual into the available tools.[3]

Let us explain in detail how to use the ANOVA-based tools referenced above. The tool receives as input the following parameters:

- α, the probability of Type I error (i.e., the probability of detecting a difference that is not real).
- β, the probability of Type II error (i.e., the probability of missing a real difference), which allows to achieve $100(1 - \beta)$ statistical power.
- m, the number of documents that will be compared using the ANOVA model.
- min_D, the minimum difference in the scores we want to detect; this parameter ensures that whenever the true difference between the most and less-relevant document among a group of m, as measured by a specific judgment scale, is equal to or greater than min_D, the goal is to ensure a statistical power of $100(1 - \beta)$.
- $\hat{\sigma}^2$, the variance estimates for the judgments; how to obtain such estimation is deeply discussed in Reference [50, Section 6.7].

The tools supports combinations of α and β where α is either 0.01 or 0.05, and β is 0.10, 0.20, 0.30, 0.40, or 0.50; thus, it is possible to follow Cohen's convention [11, 18] of $\alpha = 0.05$ and $\beta = 0.20$. By inputting m, min_D, $\hat{\sigma}^2$ into either a specific Excel sheet or using those values as parameters for the Python code for a combination of α and β, the recommended worker set size can be easily obtained. For example, if you want to ensure 80% statistical power when comparing 10 documents

[2] The same set of assumptions are being made in related work [48–50].
[3] See either the Excel file available at http://sakailab.com/download/ or the Python file available at https://github.com/KevinRoitero/set-size-estimation.

using one-way ANOVA at $\alpha = 0.05$ with an estimated population variance of $\hat{\sigma}^2 = 0.1$, and a true difference of 0.1 or more between the most and least relevant document, then use m, min_D, $\hat{\sigma}^2$ equal to 10, 0.10, and 0.10, respectively. The minimum recommended worker set size is than computed, and it satisfies the required statistical power under the specified conditions.

Let us now briefly discuss how the tools obtain the required sample size for one-way ANOVA. Let us recall that ANOVA statistics are given by

$$F_0 = \frac{V_A}{V_{E1}}; V_A = \frac{S_A}{\sigma_A}; V_{E1} = \frac{S_{E1}}{\sigma_{E1}}.$$

The test statistic F_0 essentially compares the between-group variation to the within-group variation. The probability of rejecting the null hypothesis H_0 in one-way ANOVA is given by

$$1 - P\left(F_0 \leq F_{inv}(\sigma_A; \sigma_{E1}; \alpha)\right).$$

If H_0 is true, then the probability of rejecting H_0 is exactly α. However, if H_0 is false, then such probability represents the statistical power $(1 - \beta)$, and $F_0 \simeq F'(\sigma_A; \sigma_{E1}; \lambda)$, where $\lambda = n\Delta$ and Δ measures the sum of squared document effects in variance units, that is, $\Delta = \sum_{i=1}^{m}(\mu_i - \mu)^2/\sigma^2$. We can then compute the statistical power by estimating $1 - \beta$ by leveraging the formulas detailed above; the full derivation proof can be found in Reference [50, Sections 6.4.2 and 1.3.3]. Since, in ANOVA, we have $m \geq 2$ documents that all affect the effect size Δ, to ensure the required statistical power, we need to consider the range of m population means, thus the true difference between the most- and least-relevant document among the set of m. We require $100(1 - \beta)$ statistical power when $D \geq$ min_D; to this end, we can define $min\Delta = \text{min\_D}^2/2\sigma^2$, and it can be proven (see Reference [50, Section 6.4.2]) that $\Delta \geq min\Delta$, that is, $min\Delta$ is a lower bound for Δ. Hence, we can derive that the required number of workers as

$$n \approx \frac{\lambda}{min\Delta} = \frac{2\sigma^2\lambda}{\text{min\_D}^2};$$

min_D is provided in input being the minimum detectable difference, σ^2 can be estimated, and λ can be approximated using properties of the F distribution (see Reference [50, Equation (6.39)]). The tools detailed above automatize the process of finding n.

Throughout our experiments, we follow the Cohen's five-eighty convention [11, 18] and we thus set the parameters as follows: $\sigma^2 = V_{E1}$; $\alpha = 0.05$ and $\beta = 0.2$; min_D = 0.05, where we recall that min_D represents the minimum difference in the scores we want to detect with $100(1 - \beta)\%$ statistical power; and diff_m = 2, where diff_m represents number of documents we want to be able to distinguish in a statistically significant way.[4]

Thus, for the sake of simplicity and readability, in the remainder of the article, we simply say "achieving statistical power" although it would be more precise to say "achieving 80% statistical power" (due to Cohen's five-eighty convention; see above).

In the case where the m-by-n document-by-worker matrix containing the scores is not available (think, for example, of a new crowdsourcing experiment for which no data has yet been collected), the σ^2 can be estimated from multiple datasets. In more detail, if we have C as the number of input matrices with the assessments expressed in the same scale, and we denote with n_C the number of workers of the Cth matrix and with σ_C^2 the σ^2 of the Cth matrix, then σ^2 can be computed as

[4] diff_m is referred to using the letter m both in References [48–50] and in the Excel files linked above. In this article, we use a different notation to distinguish the parameter diff_m from the parameter m denoting the dimensions of the document-by-worker matrix.

Table 1. Experimental Setting (Adapted from Reference [46])

Topic id	402	403	405	407	408	410	415	416	418
No. docs	278	111	214	212	188	212	179	174	243
No. HITs	460	182	354	350	310	350	295	287	402
Topic id	420	421	427	428	431	440	442	445	448
No. docs	164	342	195	253	203	264	408	210	419
No. HITs	270	567	322	419	335	437	677	347	695

The total number of documents and HITs is 4, 269 and 7, 059.

follows:

$$\sigma^2 = \frac{\sum_C (n_C - 1)\, \sigma_C^2}{\sum_C (n_C - 1)}.$$

Note that using the ANOVA-based set size design with diff_m = 2 is equivalent to using the set size design for two-sample t-tests, as one-way ANOVA with diff_m = 2 is exactly equivalent to a two-sample t-test. For further details see Sakai [50, pp. 124–125].

5 EXPERIMENTAL VALIDATION

5.1 Data

For the experimental validation, we considered the reassessment using crowdsourcing of 18 topics from the TREC-8 Ad-Hoc collection [25] by Roitero et al. [46] and of 36 topics from the *WebCrowd25k* dataset from McDonnell et al. [34] and Kutlu et al. [31].

In more detail, we first considered the datasets detailed by Roitero et al. [46], which have been created as a result of reassessing the documents from the TREC-8 Ad-Hoc collection using different scales. Their crowdsourcing tasks have been published using the Crowd-Flower platform, which has been also known at later times by the names Figure Eight and Appen. Each of the workers recruited was required to assess 10 documents. Two out of 10 documents assessed have been used to check the quality of the work performed by the workers, and other quality checks have been embedded into the crowdsourcing task to ensure the high quality of the data collected. Each worker was asked to provide a relevance assessment using a given relevance scale. The documents have been assessed by publishing a crowdsourcing task for each relevance scale used, this being the only difference within each task's design. The documents distributed in each single and self-contained work unit performed by an individual crowd worker did not change, nor did their ordering; in the following, we thus refer to a single work unit using the term **HIT (Human Intelligence Task)** [39]. Each document has been evaluated by 10 distinct workers, and the number of documents and HITs used for each topic is further detailed in Table 1. The relevance we considered are the following:

- S_2 [46]: binary scale with values "Not Relevant" and "Relevant," represented using the numbers 0 and 1. It is the one used originally by NIST experts to assess the relevance of documents; we denote the set of documents assessed by experts with $TREC_2$.[5]
- S_4 [46]: four level ordinal scale with values "Not Relevant," "Marginally Relevant," "Relevant," and "Highly Relevant," represented using the numbers in range $[0, 3]$. It is the one used by Sormunen [57].
- S_{100} [44, 46]: interval scale with values in the range $[0, 100]$.

[5] At the risk of sounding too pedantic, we anyway remark that the subscript in $TREC_2$ represents the binary scale used, and not a TREC edition (for which we use a dash, like TREC-8).

Then, we used the *WebCrowd25k* dataset [31, 34] made by 25,099 information retrieval relevance judgments collected using the Amazon Mechanical Turk platform. The dataset is composed of 50 different topics from the TREC 2014 Web Track[6] for which the authors have selected 100 ClueWeb12 documents[7] that have been assessed by five different workers using a four-level judgment scale. We found by comparing the two datasets that many documents in *WebCrowd25k* are assessed by less than five workers (even by one or two workers), while the number of judgments for each document in Roitero et al. [46] datasets is equally distributed, being always 10 for all the documents. Therefore, in this work, we consider the subset of 36 topics (out of 50) of the *WebCrowd25k* dataset where every document has been assessed by at least two workers. For all of the other topics, we did not have enough judgments to be able to compute the set size estimation, thus, we decided to discard them.

We choose to avoid using two publicly available datasets to validate our experimental setting due to the following reasons: The S_∞ dataset [33, 46, 58] assessments are expressed using a ratio scale with values in the range $]0, \infty[$. Such data can not be considered, because the collected crowdsourced scores need to be normalized before using them (see Reference [33, Section 4.3]). Thus, we can not use them for our experiments, because the normalization function uses the whole document-by-worker matrix for a given topic. Furthermore, the $]0, \infty[$ scale employed in the S_∞ dataset is not suitable for the statistical tests employed in this work, as the judgments are log-normally distributed (see Reference [33, Section 4.2]) [28]. Furthermore, we do not consider the assessments collected by Yang et al. [62], as they focus on pairwise judgments (thus using a different experimental setting when compared to the other datasets used in this work) and assess a small subset of the documents already assessed by Maddalena et al. [33] and Roitero et al. [44].

Summarizing, we use two publicly available datasets, the ones provided by Roitero et al. [46] (S_2, S_4, and S_{100}) and the one provided by McDonnell et al. [34] and Kutlu et al. [31] (*WebCrowd25k*). We refer to their works for the ethics approval on the data published by them and used in this work.

5.2 RQ1: Estimation of the Number of Workers

5.2.1 Aims and Settings. This experiment aims estimating, for each topic separately, the number of workers recommended to achieve enough statistical power to distinguish two documents. For each topic, we build the document-by-worker matrix and we set the parameters as described in Section 4 for the estimation methodology employed.

5.2.2 Results. Let us first focus on Figure 1 top row. We show on the x-axis the number of actual workers and on the y-axis the number of workers recommended according to our methodology such that we have enough statistical power to be able to distinguish in a statistically significant way two documents of the topic. A point that lays on the upper-left part of the plot, above the dashed gray line, means that the number of workers for such a topic is smaller than what is recommended (i.e., such topic, or rather its documents, do not have enough statistical power). Vice versa, a topic that lays on the lower-right part of the plot means that the number of workers for such a topic is higher than what is recommended (i.e., such topic, or rather its documents, have enough statistical power). As we can see, many topics (about 50% of them for all the considered datasets) have enough statistical power when considering the S_2, S_4, and S_{100} datasets, while the same does not hold when considering the *WebCrowd25k* dataset (fourth plot of the row) where there is not enough statistical power for any of the topics. This is both a positive and negative result for researchers and practitioners who plan to use the collected crowdsourced relevance assessments in a test collection

[6] http://www-personal.umich.edu/~kevynct/trec-web-2014/.
[7] https://lemurproject.org/clueweb12/.

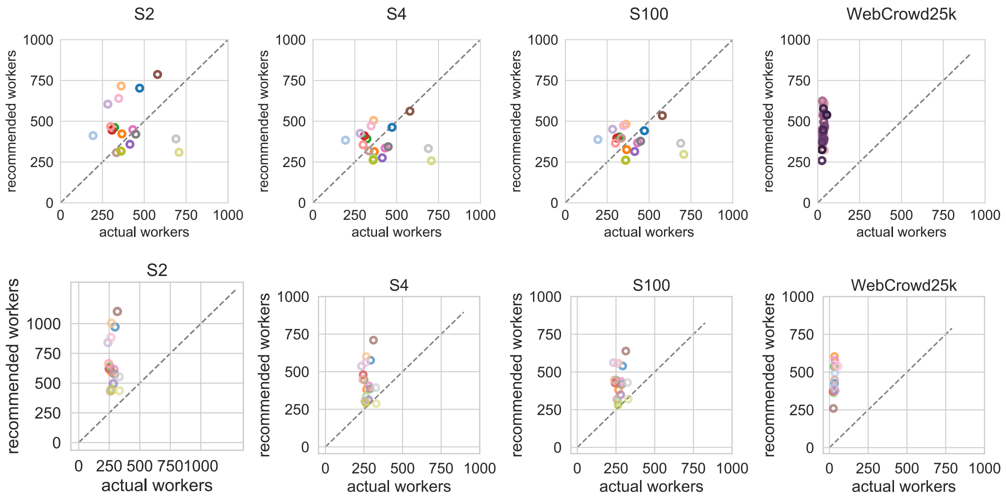

Fig. 1. Actual (x-axis) and recommended (y-axis) number of workers for each considered dataset (top row) and when bootstrapping the number of documents and workers for S_2, S_4, and S_{100} while considering only the subset of topics with at least four judgments per document on *WebCrowd25k*. Each dot is a topic. The $x = y$ line is in dashed gray.

setting. In fact, for half of the topics in three out of four datasets, the relevance assessments are statistically meaningful. That is, if the relevance levels of documents A and B truly differ, then there is a high chance that that is reflected in the assessments. However, for half of the topics, we observe the opposite. Ideally, to ensure a consistent evaluation across the whole test collection, we would need all topics to have enough statistical power. We also observe a high topic variability and that there is a rather high consistency across datasets: The three plots look overall similar and the point corresponding to each topic tends to fall roughly in the same region of the plot across different scales.

By looking again at Figure 1, we see that the results for the *WebCrowd25k* dataset seem to be different from what has been previously observed for S_2, S_4, S_{100}. Since the two sets of judgments have been collected on different collections with different experimental settings, we perform an additional set of experiments to make such sets of judgments as comparable as possible. In more detail, we use the bootstrap technique to random sample 100 documents evaluated from four workers in the S_2, S_4, and S_{100} datasets; in this way, we mimic the setting used to collect judgments for the *WebCrowd25k* dataset, thus allowing us to directly compare the results. We repeat the sampling process 100 times. Moreover, to further improve the quality of the comparison, we remove from the *WebCrowd25k* dataset the topics containing documents with less than four judgments; by doing so, we are left with 14 topics. The results of these experiments are reported by Figure 1, bottom row, thus showing different situation from the top one. In fact, we notice that the topics for the S_2, S_4, and S_{100} dataset drift towards the left-upper part of the plot, thus not having enough statistical power. This behavior confirms that the results shown in Figure 1, top row, generalize across collections. In fact, as we can see considering the bottom row, we have found evidence of a similar behavior (the majority of the topics lay above the $x = y$ line when performing bootstrap for all the considered collections and thus a disjointed set of topics, collected in different years and TREC tracks).

To complete our analyses concerning RQ1, we now address some notable topic features to further investigate the topic variability, as reported in Figure 2. To this aim, we consider three topic

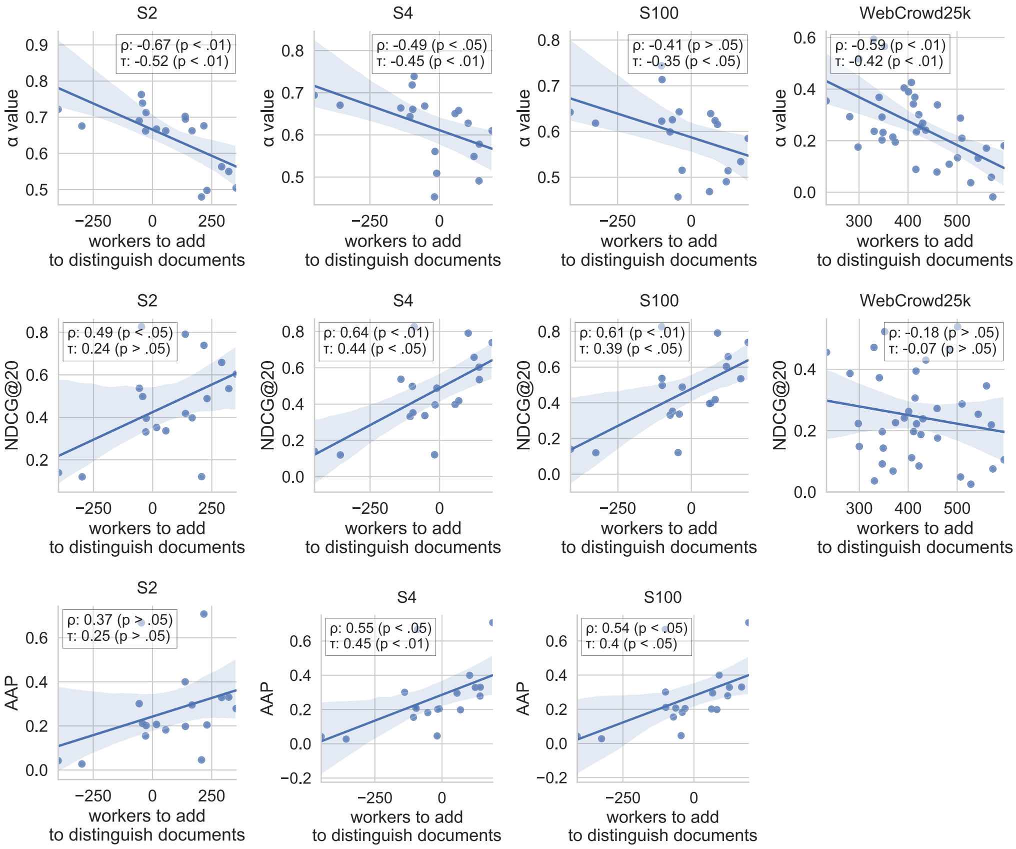

Fig. 2. Top row: correlation between the number of recommended workers to add (x-axis) and Krippendorff's α (y-axis) for S_2, S_4, and S_{100}. Bottom row: correlation between the number of recommended workers to add (x-axis) (x-axis) and NDCG@20 (y-axis) for all the datasets. Third row: correlation between the number of recommended workers to add (x-axis) (x-axis) and AAP (y-axis). We do not compute this correlation for the *WebCrowd25k* dataset, since the collection qrels contained duplicate documents. Each dot is a topic. Each plot is annotated with correlation values for Kendall's τ and Pearson's ρ.

features. In the top row, we show the correlation between the number of workers needed to distinguish two documents and the worker agreement, and in the subsequent two rows of the figure, we check the correlation between the official metrics employed in the TREC collection against the number of workers to add to distinguish two documents. Therefore, in Figure 2, middle row, we show the correlation between the recommended number of workers needed to achieve statistical power and the average **Normalized Discounted Cumulative Gain (NDCG@20)**, while in the bottom row, we show the correlation with the topic difficulty (as perceived by retrieval systems), thus measured by taking the average system effectiveness score for each topic over the systems. As a result, we use the **Average Average Precision (AAP)** [35, 45] for the S_2, S_4, and S_{100} datasets, and for all of the datasets.

Let us first consider Figure 2, bottom row, which shows the correlation between the number of workers to add to distinguish documents (x-axis) and the AAP measure (y-axis) for S_2, S_4, and S_{100} datasets. We do not compute this correlation for the *WebCrowd25k* dataset, since the collection's qrels contained duplicate documents, and trec_eval (i.e., the official evaluation software) could not

Table 2. ANOVA Table for Effect Size

	F	**PR(>F)**	η^2	ω^2	**Effect**
Dataset	12.055	0.000	0.047	0.043	small effect
Document relevance (TREC_2)	43.370	0.000	0.279	0.272	large effect
Topic	4.883	0.000	0.333	0.264	large effect
No. documents	2.313	0.130	0.003	0.002	no effect
No. workers	6.259	0.013	0.008	0.007	no effect
Topic Difficulty (NDCG@20)	3.639	0.058	0.005	0.003	no effect
Worker Agreement (Krippendorff's α)	1.581	0.210	0.002	0.001	no effect

compute AP scores for such collection. From the plots, we can see evidence of a positive correlation between the recommended number of workers to add to the set and the topic difficulty measured with AAP. Therefore, when the topic has generally higher effectiveness over the systems participating in the corresponding TREC track (i.e., the topic is more difficult), the number of workers recommended to achieve 80% statistical power is higher. Particularly, this correlation is statistically significant for the S_4 and S_{100} datasets. This behavior is consistent for all the aforementioned datasets when measuring the correlation between the recommended number of workers to add to the set and the NDCG@20 measure, as can be seen in Figure 2, middle row. However, this does not hold when considering the *WebCrowd25k* dataset, which shows a weak and not significant negative correlation. The different behavior of the S_2, S_4, and S_{100} datasets with respect to *WebCrowd25k* could depend by the different underlying collection (i.e., TREC-8 for S_2, S_4, and S_{100}, and the 2014 NIST TREC web track for *WebCrowd25k*); in fact, the two collections employed different settings, with different relevance scales and a different evaluation methodology.

We now turn to discuss another topic features, namely, the worker agreement. Figure 2, top row, shows the correlation between the recommended number of workers to add to distinguish documents and the per-topic inter-worker agreement, computed using Krippendorff's α [30]. In the figure, each plot is annotated with the correspondent correlation values both for Kendall's τ and Pearson's ρ. When inspecting the plots, we see that a negative correlation holds between the number of workers to add and workers' agreement (i.e., for topics with a lower agreement, a higher number of workers is recommended). The observed correlation is statistically significant across all the datasets, exception made for Pearson's ρ in the S_{100} dataset. In practice, this means that the more agreement we observe on the collection topics, the fewer workers we need to add to be able to distinguish in a statistically significant way documents. In other words, workers' agreement acts as a proxy to estimate the total number of workers needed to achieve statistical power.

Finally, we compute an ANOVA analysis using the variables listed in Table 2 to draw a general picture and investigate how the different factors and dimensions affect the number of recommended workers. We compute the ω^2 index [38] as representative of the size of effect, using 0.01, 0.06, and 0.14 as thresholds to measure effect size[8] [23]. As we can see from Table 2, we find that the dataset has a small effect on the size of the recommended workers to distinguish two documents in a statistically significant way, while the TREC_2 relevance and the topic have a large effect on the size of such set. All of the other considered dimensions such as the number of documents and workers, the NDCG@20, and the Krippendorff's α for each topic have no effect. We perform an in-depth analysis of those factors in the following sections of the article. Overall, this analysis shows that our results can be generalized across collections particularly when considering some topic

[8] Similar analyses have been proposed in the literature to understand the contributions of different system components when compared to quality indicators [19, 21, 22, 42, 63].

features such as the workers' agreement computed using the Krippendorff's α. In all of the other topic features that we have considered, we need more data to draw generic conclusions. Moreover, the major effect on the estimation of the recommended number of workers is derived from the specific topic and document relevance; thus, the estimation of recommended number of workers should be performed per topic, considering the expert relevance of the documents for which we are collecting relevance judgments (if known in advance). Furthermore, since the specific dataset has a small effect, researchers and practitioners can use the estimation on one collection (if either the same set of topics or documents has been used) to obtain a reliable estimation for their experiment.

5.2.3 Take-home Messages. We can draw different remarks from the results described in Section 5.2.2. Overall, there is a high topic variability, meaning that different topics behave differently; thus, when designing a crowdsourcing task to collect relevance assessments, each topic should be treated separately, and the parameters that are suitable for one topic are not suitable for another one. As a result, researchers cannot estimate the residual variance of the worker-by-document matrix by considering one topic and then applying the same result to other topics. This issue is also studied in depth in the following sections.

While we can observe some minor differences between the used datasets (probably due to the different settings used to build each of these collections, as suggested by Figure 1, middle row), the ANOVA analysis shows that such differences have a small effect on the estimation of the number of workers needed to achieve statistical power. Thus, the set size estimation on one collection (if either the same set of topics or documents has been used) can be used to obtain a reliable estimation for the experiment.

We also note that there is little to no scale variability for the effects measured on the different topics: The same topic tends to behave similarly across different datasets (i.e., S_2, S_4, S_{100}). This means that the data collected using a certain relevance scale can be used to reliably estimate the matrix residual variance and thus the recommended number of documents and workers needed to achieve statistical power when using a different scale. Furthermore, the judgments collected using a scale can be transformed to another scale before the set size estimation, since the scale transformation has a negligible effect.

Last, we note that for nearly half of the test collection topics and particularly for the *WebCrowd25k* collection, we would need additional workers to achieve statistical power and thus be able to reliably distinguish document scores. This is a rather concerning result: If a researcher or a practitioner chooses a too small amount of crowd relevance assessments for the topics used in the evaluation, then the collected judgments can lead to several type II errors.

5.3 RQ2: Impact of Document Relevance

5.3.1 Aims and Settings. This experiment aims to analyze the impact of the document labels provided by TREC experts to understand whether document relevance has an impact on the worker set size estimation. Table 2 already shows that the TREC_2 relevance has an impact on the worker set size estimation. As for the previous section, we set the experiment's parameters as described in Section 4.

5.3.2 Results. Figure 3 shows, similarly to Figure 1, the number of actual and recommended workers needed to have enough statistical power to distinguish documents across all the considered datasets. The data is broken down along the TREC_2 relevance levels.

As we saw in Figure 1, in Figure 3 the number of topics lying below the dashed line is roughly consistent across the first three considered scales. When considering TREC_2 relevant documents, the majority of topics lay above the dashed line, even if there is a little dataset effect: 14 out of 18

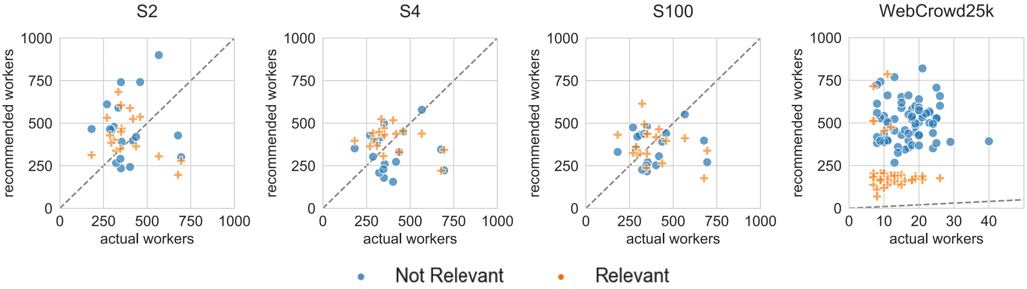

Fig. 3. Actual (x-axis) and recommended (y-axis) number of workers. Each dot is a topic. The $x = y$ line is in dashed gray. Breakdown on $TREC_2$ (note that each topic appears twice).

for S_2, 12 for S_4, and 11 for S_{100}. Conversely, when considering $TREC_2$ not-relevant documents, less than half of the topics lay above the dashed line; more precisely, 10 out of 18 for S_2, 8 for S_4, and 7 for S_{100}. This does not hold for the *WebCrowd25k* dataset, where for both the relevance values the documents lay above the dashed line (i.e., for each relevance value and for each considered topic, we do not have enough statistical power). Therefore, only for the S_4 and S_{100} datasets when considering not-relevant documents the number of actual workers is sufficient to achieve statistical power for the majority of topics, while this does not hold for relevant documents. We conclude that in the majority of the topics for all the datasets, we do not have statistical power when considering the $TREC_2$ breakdown and that, in general, there is more statistical power for the not-relevant documents.

Then, we investigate whether such a behavior is influenced by the different levels of agreement that $TREC_2$ relevant documents may have when compared with $TREC_2$ not-relevant ones. Thus, we compute the per-topic inter-worker agreement using Krippendorff's α for both the relevant and not-relevant documents, and we compare it with the number of documents for the topic. Checco et al. [6] provided evidence that α depends on the number of documents used to compute the agreement within a crowdsourcing setting. To solve this issue, we compute the agreement in a twofold way. First, we compute α for each $TREC_2$ relevant and not-relevant document. Second, we use the bootstrap technique [26] to remove the effect of documents' number and we sample with replacement the same number of relevant and not-relevant documents (the maximum number available) before computing α. We perform the sampling 1,000 times to avoid possible biases. Note that we cannot perform bootstrap for the *WebCrowd25k* dataset, since, for many topics, we do not have enough documents with a given $TREC_2$ relevance evaluated by a sufficient number of workers.

Figure 4 reports the results. By looking at Figure 4, top row, we see that $TREC_2$ relevant and not-relevant documents do not show any significant correlation with the Krippendorff's α for all the considered datasets. This result is further confirmed by looking at the bottom row of Figure 4, which shows the same correlations computed using the bootstrap technique. Therefore, the peculiar behavior of the $TREC_2$ relevant documents shown by Figure 3 cannot be explained by the different levels of worker agreement between $TREC_2$ relevant and not-relevant documents.

5.3.3 Take-home Messages. We can draw different remarks from the results described in Section 5.2.2. The labels assigned to the documents by $TREC_2$ experts have no or negligible effect when estimating the number of workers needed to distinguish documents in a statistically significant way, in the sense that while the relevant and not-relevant documents show a different behavior (as confirmed by Table 2), both sets of documents lead to a similar worker set size estimation and, more importantly, they are both under-powered in existing datasets.

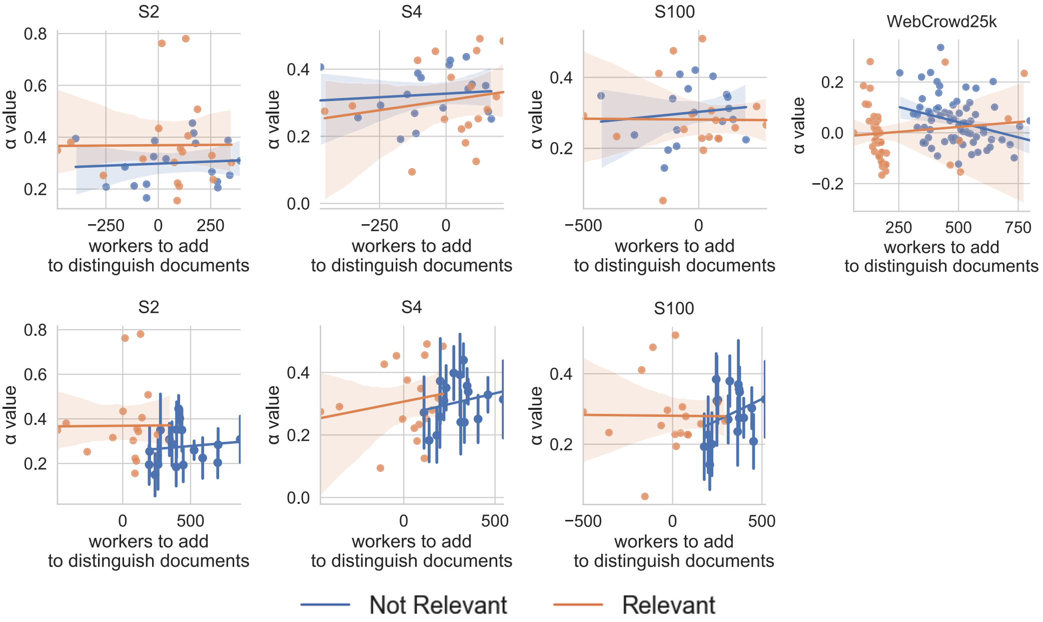

Fig. 4. Number of recommended workers to add (x-axis) with a breakdown on TREC$_2$, and Krippendorff's α (y-axis) on all the datasets (top row), or when bootstrapping the number of not-relevant documents (bottom row) for S$_2$, S$_4$, and S$_{100}$ datasets.

The results obtained suggest some practical guidelines to researchers or practitioners who aim to design a crowdsourcing task to collect relevance assessments for a set of documents. They should either use a mixture of (presumed) relevant and not-relevant documents or take the (presumed) relevance of those documents into account when estimating the worker set size recommended to achieve statistical power when analyzing workers' differences. Furthermore, they should not use the estimation done over a set of not-relevant documents to estimate the worker set size needed for a set of relevant documents and vice versa.

5.4 RQ3: Impact of Workers' Features

5.4.1 Aims and Settings. We now investigate what happens during relevance assessments collection and how crowd workers' features such as arrival time, agreement, and quality affect the estimation of the number of workers recommended to achieve statistical power and distinguish documents. In more detail, we investigate the impact of workers' arrival time and how their assessment quality impacts the overall worker set size estimation. We describe our analyses focusing on the S$_{100}$ scale dataset, since the results are very similar to the ones obtained for S$_2$ and S$_4$. *WebCrowd25k* is also omitted from this kind of analysis due to the structure of the dataset. Since for many topics there are too few workers judging a document, the results are not statistically significant. Therefore, we could not compare results across different workers, since there is not a fixed setting across all the datasets, and the number of documents evaluated by each worker has high variability. Note that, given that arrival time can not be controlled (or at least, not in a simple way), we conjecture that such a feature will not have any impact on the estimation of the number of workers recommended to achieve statistical power and, indeed, the presence of an effect might suggest the presence of bias in the data. Nevertheless, we believe it is important to verify the conjecture on real data.

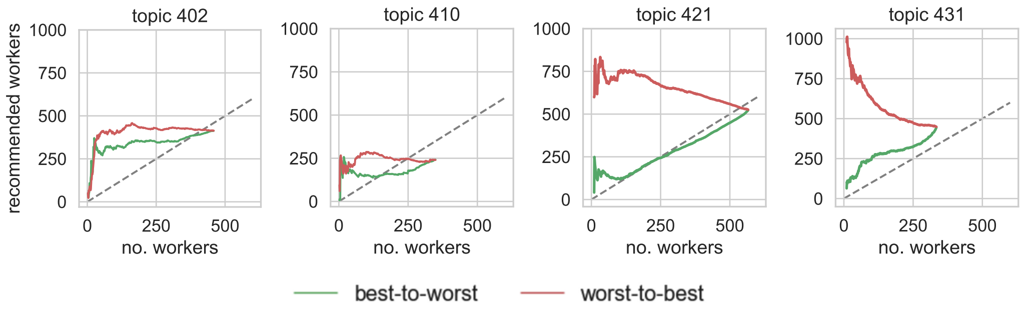

Fig. 5. Number of recommended workers (y-axis) for the cardinality of the workers set considered (x-axis) sorted according to quality (first two plots) and agreement (third and fourth plot). The $x = y$ line is in dashed gray. Sample of topics from the S_{100} dataset.

To investigate the impact of worker's arrival time, we sort the workers according to their arrival timestamp recorded by the crowdsourcing platform in three ways: (i) from the first one to the last one; (ii) from the last one to the first one; (iii) by randomly sampling them and performing bootstrap, as a baseline. We average the results over 1,000 samples for each sampling size to avoid sources of bias. Then, we measure the impact of each sorting approach on the number of workers recommended to achieve statistical power for each cardinality of the workers set considered.

The quality of the work performed during a crowdsourcing experiment is usually not known *a priori*. Since crowdsourcing experiments are often published in the form of an open call [27], workers with different backgrounds, knowledge, and level of effectiveness contribute to the experiment's result [15, 16]. It is thus interesting to investigate the impact of worker quality on the estimation of the number of workers recommended to achieve statistical power and distinguish documents. To this end, we measure the quality of the relevance assessments provided by each worker by computing the accuracy score of the worker within the set of documents s/he assesses during the crowdsourcing experiment. Then, we sort the whole workers set to simulate their arrival time in either increasing or decreasing order according to their accuracy score. Last, we compute the recommended number of workers needed to achieve statistical power to distinguish documents for each worker's set cardinality.

We also investigate the impact of worker agreement. To compute the contribution of each worker concerning his/her agreement, we follow the same procedure adopted by Sakai et al. for the same purpose (see Reference [53, Table 1, p. 12]): (i) we compute Krippendorff's α on the whole topic; (ii) we remove one worker at a time and we recompute Krippendorff's α on the whole topic. The contribution of each worker to the overall agreement of a topic is then computed by considering the absolute value of the difference between the two α values. Then, we sort the workers according to their contribution to the overall agreement and we compute, for each considered worker set cardinality, the recommended number of workers we need to add to achieve statistical power and distinguish documents.

5.4.2 Results. The results are summarized in Figure 5 for a subset of the S_{100} dataset, since the results are consistent across all the considered ones. In the figure, we omit the results when sorting the workers by their arrival time, since we found that such a worker feature leads to similar results for all the approaches described in Section 5.4.1. This suggests that workers' arrival time does not affect the number of workers to add to achieve statistical power when distinguishing documents. Such a result holds for each topic of all the considered datasets.

Figure 5's first two plots report the results concerning the impact of workers' quality. We report only a small subset of topics in this figure, since they all behave similarly; the full set of plots is made available in the repository associated with this article. The x-axis shows the cardinality of the workers' sets, while the y-axis shows the recommended number of workers to add to achieve statistical power. The workers within each set are sorted from the lowest-to-highest accuracy score (red line) or vice versa (green line). The number of workers to add to achieve statistical power is generally higher when sorting them according to their accuracy scores in ascending order, as expected.

Finally, the third and fourth plots of Figure 5 report the results concerning the impact of workers' agreement. The x-axis shows the cardinality of the workers' sets, while the y-axis shows the recommended number of workers to add to achieve statistical power. The workers within each set are sorted from the lowest to highest contribution to the overall agreement (red line) or vice versa (green line). There is a noticeable gap between the two strategies, meaning that when there is a high agreement between workers the number of workers to add to achieve statistical power is quite lower and vice versa.

5.4.3 Take-home Messages. We can draw different remarks from the results detailed in Section 5.4.2. Workers' arrival time does not affect the number of workers to add to achieve statistical power. Researchers and practitioners who want to collect relevance assessments can use workers' judgment without worrying about workers' arrival time. Furthermore, they should prioritize assessments provided by high-quality workers, since it allows them to reduce the number of workers to add. Last, researchers and practitioners should prioritize as well assessments provided by workers with a high inter-annotator agreement. This recommendation has a practical impact, since inter-annotator agreement can be computed without requiring a ground truth for document relevance provided by experts. Before moving on, let us make some remarks on the prioritization of high-quality and high-agreement workers. While at a first glance, it might seem obvious that high-quality and/or high-agreement workers should be prioritized, we believe that this is not always true. This is more evident in the high-agreement workers' case. In fact, it might happen that a set of workers express a set of judgments over a document that has the very same value, but such a value is different from the real relevance value for such a document. This case leads to the maximum agreement but low quality and those workers should not be prioritized.

5.5 RQ4: Upper and Lower Bound for Worker Set Size Estimation

5.5.1 Aims and Settings. This experiment investigates one of the limitations of the approaches that deal with workers' features proposed in the previous sections to estimate the number of recommended workers to achieve statistical power. Such approaches propose only a rough estimation of the optimal result (i.e., maximum statistical power with the minimum number of workers) that we can achieve with a subset of workers. Instead, we now focus on a more accurate estimation.

We compute for each cardinality of the worker set for a document-by-worker matrix of a topic the workers subset that provides the most statistical power (i.e., minimizes the number of recommended additional workers). We call the resulting set computed for each cardinality the *Best* series. In other words, the Best series is a series obtained considering for each cardinality the subset of workers that makes more evident the difference between two documents. Symmetrically, we study which is the worker subset that provides the least statistical power (i.e., maximizes the number of recommended additional workers). We call the resulting set the *Worst* series. In other words, this time we generate the Worst series by choosing for each cardinality the subset of workers that makes two documents less easy to distinguish. Last, we compute for each cardinality of the worker set the average number of recommended workers to add by random sampling a given amount of workers for 10,000 times. We call the resulting series as *Average* and serve a double purpose: It acts as a baseline for the Best and Worst series and shows how many workers are needed on average to reliably estimate the recommended number of workers for a document-by-worker matrix.

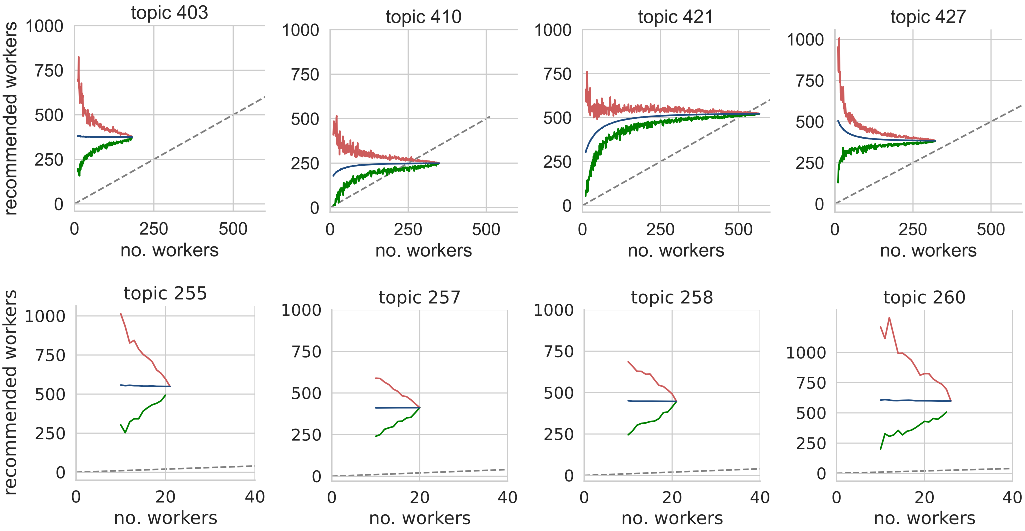

Fig. 6. Number of workers (x-axis) and recommended workers (y-axis) for the Best, Worst, and Average series for a few topics. Top row is for S_{100} and bottom row is *WebCrowd25k*.

We propose an approach based on **Evolutionary Algorithms (EAs)** to perform this experiment due to the high dimensionality of the search space. The approach proposed has been successfully applied to a similar problem that concerns estimating and identifying the best correlation we can achieve with a subset of topics [3, 24, 43, 47]. Our approach employs the NSGA-II [13] algorithm. The software needed to replicate our experiments is publicly available at https://github.com/Miccighel/Crowd-Size-Gen-EA together with a detailed description of the algorithm's parameters and settings.

5.5.2 Results. Figure 6 shows the estimated (x-axis) and recommended (y-axis) number of workers to add needed to distinguish documents in a statistically significant way for a sample of topics from the S_{100} (top row) and the *WebCrowd25k* (bottom row) datasets. We show only a subset of topics, since all the others behave similarly. The plots for all the topics for each dataset are made available via a public repository *The plots for all the topics for each dataset are made available in the public repository linked above.* As shown by the figure, there is a gap between the Best and Worst series for all the topics. Both series increase. Thus, the more workers we consider, the more the number of workers to add converges towards its real value computed using the full document-by-worker matrix. The fluctuations in the Best and Worst series are due to the parameters provided to the EA used and can be smoothed by increasing the number of iterations performed [47].

The Average series quickly converges towards the real number of recommended workers to add needed to achieve statistical power. This is particularly evident for the *WebCrowd25k* dataset, where the Average series looks like a straight horizontal line due to the lower number of workers per topic on the dataset. This result has practical implications, since it can be used to estimate beforehand such a number; we discuss this result in depth in Section 5.6.1.

5.5.3 Take-home Messages. We can draw two remarks from the results described in Section 5.5.2. Researchers and practitioners ideally want to estimate the recommended number of workers to add to achieve statistical power using a subset of workers; however, the choice of a certain worker subset has an impact. Furthermore, there is a huge gap between the subsets of

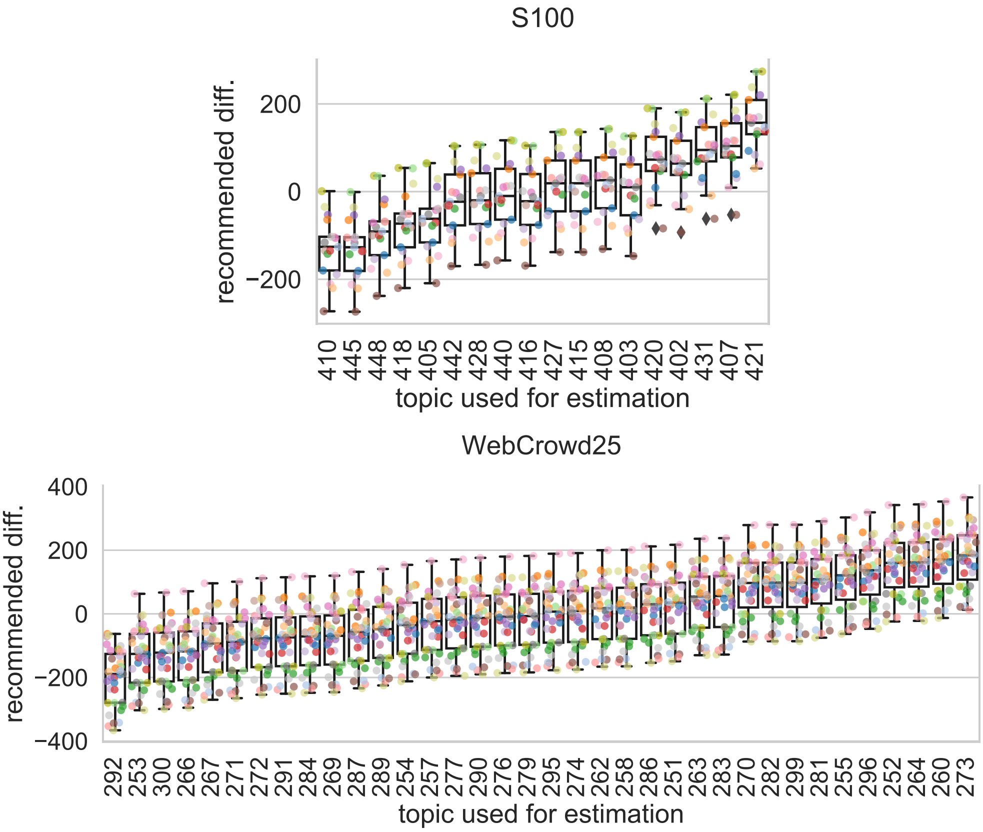

Fig. 7. Estimation topic (x-axis) and recommended number of workers difference between estimation and estimated topic (y-axis) for S_{100} (top row) and *WebCrowd25k* (bottom row). Box-plots are sorted by increasing median value.

workers that maximize/minimize the recommended number of workers to add, and a random sample of workers is the best choice in practice.

5.6 RQ5: Practical Estimation Strategies

5.6.1 Aims and Settings. We now identify practical strategies that can be used to reliably estimate the recommended number of workers to add to achieve statistical power to distinguish documents before performing a crowdsourcing experiment. We try to use the residual variance needed to compute such a number of workers for a single topic as a proxy to estimate the residual variance and compute such a number for the remaining topics. Then, we analyze and leverage the predictive power of the Average series described in Section 5.5.2. As before, we report the results only for S_{100} and *WebCrowd25k*, since the others show similar patterns (see the public repository for all the other plots).

5.6.2 Results. Figure 7 shows the topics used to estimate the recommended number of workers to add (x-axis) and the difference between the number computed using a single topic with respect to the other topics (y-axis) in the top row for S_{100}, while in the second for *WebCrowd25k*. To provide an example, we focus on the S_{100} plot, where the leftmost box-plot shows such a difference when using topic 410 to estimate the number for the remaining topics; in this case, for most of the topics, the number computed is an overestimation. Overall, there is high topic variability; certain topics

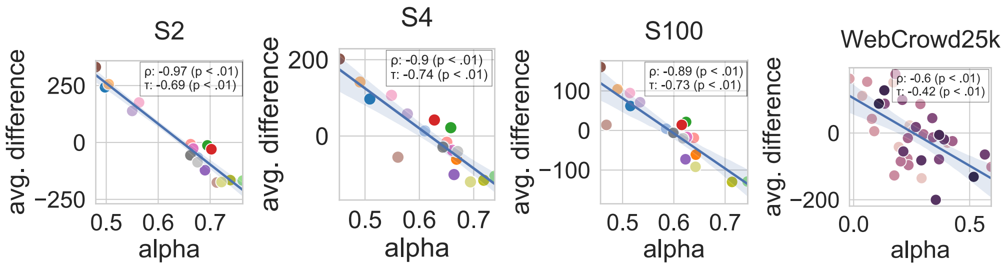

Fig. 8. Correlation between Krippendorff's α (x-axis) and average per-topic difference between estimation and estimated worker set sizes (y-axis) for all the datasets.

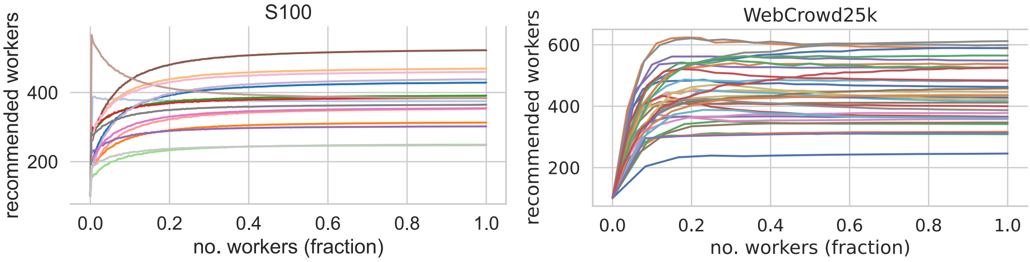

Fig. 9. Number of workers (fraction) on x-axis and recommended number of workers on y-axis for the Average series. Top row is for S_{100} and bottom row is *WebCrowd25k*.

estimate adequately the recommended number of workers to add to achieve statistical power for the remaining ones while others do not. We correlate for each topic the AAP and the Krippendorff's α with the average-per-topic difference.

Figure 8 shows the result for Krippendorff's α for all the datasets. The plot for AAP is omitted, since there is no correlation. In each dataset there is a significant negative correlation between the inter-worker agreement measured for a topic and the ability of the topic to reliably estimate the recommended number of workers to add to achieve statistical power. The 0 point lies in the middle of the y-axis; this means that a higher agreement for a topic leads to a higher overestimation of the recommended number of workers to add.

Figure 9 shows the fractions of the total number of workers available for a given topic (x-axis) and the number of recommended workers to add to achieve statistical power (y-axis) for S_{100} and *WebCrowd25k*. Each line represents a different topic. When considering a fraction of workers between the 20% and the 40% the curves become almost horizontal. This means that we can reliably estimate the recommended number of workers to add to achieve statistical power when using the full workers set. We also note that for certain topics the series converges in reverse order (i.e., they monotonically decrease instead of increase); thus, while for the majority of topics a smaller fraction of workers selected results in an underestimation, for a small fraction of them it results in an overestimation. This is particularly evident for the *WebCrowd25k* dataset (second plot), where the number of recommended workers converges faster to a specific value. We leave an in-depth analysis of the possible causes of such behavior for future work.

5.6.3 Take-home Messages. We can draw two remarks from the results described in Section 5.6.2. Researchers and practitioners who need to estimate the residual variance for an unknown topic should use the residual estimation computed using topics with a high inter-worker agreement. Moreover, they can collect a small amount of data for the unknown topic and use it to

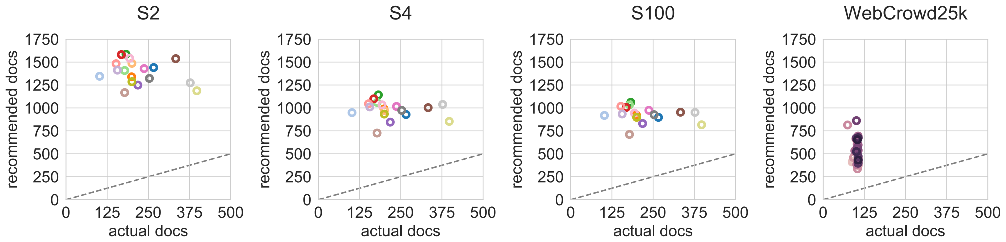

Fig. 10. Actual (x-axis) and recommended (y-axis) number of documents for each considered dataset for S_2,S_4, S_{100}, and *WebCrowd25k*. Each dot is a topic. The $x = y$ line is in dashed gray.

estimate the recommended number of workers needed to obtain statistical power for such a topic (cf. Figure 9).

6 RQ6: ESTIMATING THE NUMBER OF DOCUMENTS TO ACHIEVE STATISTICAL POWER

6.1 Aims and Settings

So far, our focus in the article has been to estimate the number of workers needed to distinguish two documents in a statistically significant way by using the document-by-worker matrix. In doing so, we have considered different topic features, such as TREC_2 relevance and features. By looking at the problem formalization as detailed in Section 4, we can notice that our proposed approach is general, and one could consider leveraging the worker-by-document matrix to being able to distinguish two workers in a statistically significant way. The worker-by-document matrix, as previously discussed in Section 1, is used to compute the number of documents recommended to distinguish in a statistically significant way two or more crowd workers; in this case, the aim is different: We want to perform some analyses such as identifying the workers with higher quality or analyzing some other worker features (e.g., identify the subset of workers with more agreement with experts, the more consistent ones). To this aim, we are considering the complementary problem as the one solved before: We want to understand how many documents the workers need to evaluate to draw statistically significant conclusions about workers thus being able to distinguish in a statistically significant way two workers. As already discussed in Section 4, the methodology does not change, as it is sufficient to consider the transposed matrix (i.e., workers-by-documents instead of documents-by-workers) and then apply the same techniques as before.

6.2 Results

As done when estimating the number of workers in Figure 1, in Figure 10, we show on the x-axis the number of actual documents and on the y-axis the number of documents recommended according to our methodology such that we have enough statistical power to be able to distinguish in a statistically significant way two workers. Again, a point that lies on the upper-left part of the plot, above the dashed gray line, means that the number of documents for such a topic is smaller than what is recommended. Vice versa, a topic that lies on the lower-right part of the plot means that the number of documents for such a topic is higher than what is recommended. By inspecting the plots of Figure 10, it can be seen that the whole set of topics is above the $x = y$ dashed line, consistently, for each of the datasets considered. Therefore, while the number of recommended documents varies a lot across the datasets (particularly, there is a huge difference between S_2 and *WebCrowd25k*), for each topic, we do not have enough documents to distinguish two workers.

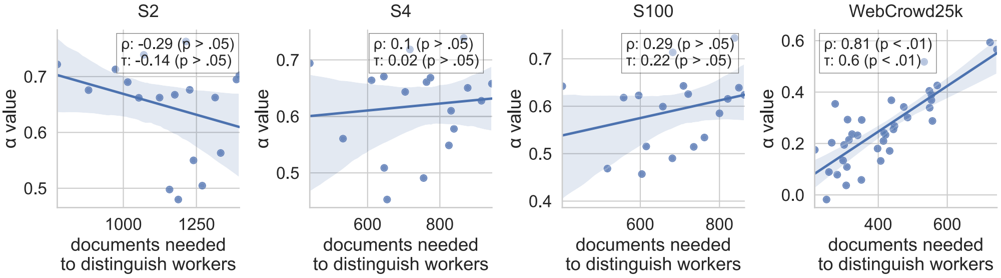

Fig. 11. Correlation between recommended number of documents (x-axis) and Krippendorff's α (y-axis). Each dot is a topic. Each plot is annotated with correlation values for Kendall's τ and Pearson's ρ.

Consistently with what we have done before, we also consider some topic features to further investigate evidence of topic variability not found in Figure 10. Therefore, as done in Figure 2 (top row), in Figure 11, we show the correlation between the number of documents needed to distinguish two workers and the worker agreement measured with Krippendorff's α for all of the datasets.

The reported results show a statistically significant correlation only when considering the *WebCrowd25k* dataset (rightmost plot), while we have some weak and not statistically significant correlation for all of the others. Moreover, we can see that the sign of the correlations is not consistent across the datasets: While the majority of them have a positive correlation (S_4, S_{100}, and particularly strong for *WebCrowd25k*), the S_2 datasets hints a weak and negative correlation.

6.3 Take-home Messages

In light of the discussion above, we can say that we do not have enough documents to distinguish two workers in a statistically significant way in each of the considered test collections.

Therefore, if, for example, the aim of the researchers and practitioners conducting the study is to identify workers' features, when designing a crowdsourcing task, a higher number of documents would be needed. Moreover, we found little to no evidence regarding the influence of topic features on the number of documents needed to distinguish workers. While this is not the main aim of the current article, this could be an interesting topic for future work. In particular, this might be leveraged by human computation-oriented work that focuses on worker features and differences between a set of workers.

7 CONCLUSIONS AND FUTURE WORK

In this article, we propose an effective methodology based on t-test and ANOVA that allows researchers and practitioners to estimate the recommended number of workers needed to meet a predefined set of statistical requirements when crowdsourcing relevance assessments. More in detail, the main findings and contributions are as follows:

- we investigate the worker set size estimation in the setting of relevance assessment for multiple collections and assessment scales;
- we found that such estimation is dependent on the topic considered, while the effect of the assessment scale used is limited;
- we found that the ground truth $TREC_2$ relevance of the documents considered have a negligible impact on the worker set size estimation;
- we found that different workers' features impact differently the worker size estimation: While arrival time has no effect, the quality and agreement of workers has a major effect,

where workers of higher quality with the higher agreement should be employed when possible;

- we provided an upper and lower bound for the worker set size estimation, and we also provided researchers and practitioners with a set of practical strategies that allow estimating the recommended number of workers to achieve the statistical power needed to distinguish documents;
- finally, we have also investigated the estimation of the number of recommended documents needed to have enough power to distinguish between two workers.

It is important to remark that the approach proposed in this article is general and not bound to any specific domain. In this work, we focused on relevance assessments for the documents of a test collection, but the methodology proposed can be applied to any set of items being judged by a crowd of workers, i.e., every time that we can build an assessor by item matrix. Thus, this article opens the path to multiple research lines. In future work, we aim to expand our findings to other domains besides test collections. Furthermore, further experimentation and diverse domains allow taking into account additional worker or document features as well as methodologies based on confidence intervals to study their impact on the estimation of the recommended number of assessors to add to achieve statistical power to distinguish items. We will also focus on estimating not only the number of workers required but also determining which workers to employ from a larger pool.

Overall, this article is a step towards a more robust and sound approach that can be applied to reliably estimate the relevance of documents using crowd workers to build sound and reliable test collections.

REFERENCES

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

[2] Tara S. Behrend, David J. Sharek, Adam W. Meade, and Eric N. Wiebe. 2011. The viability of crowdsourcing for survey research. *Behav. Res. Meth.* 43, 3 (25 Mar. 2011), 800. DOI: https://doi.org/10.3758/s13428-011-0081-0

[3] Andrea Berto, Stefano Mizzaro, and Stephen Robertson. 2013. On using fewer topics in information retrieval evaluations. In *Proceedings of the Conference on the Theory of Information Retrieval (ICTIR'13).* Association for Computing Machinery, New York, NY, 30–37. DOI: https://doi.org/10.1145/2499178.2499184

[4] Chris Buckley and Ellen M. Voorhees. 2017. Evaluating evaluation measure stability. *SIGIR Forum* 51, 2 (Aug. 2017), 235–242. DOI: https://doi.org/10.1145/3130348.3130373

[5] Ben Carterette, James Allan, and Ramesh Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'06).* ACM, New York, NY, 268–275. DOI: https://doi.org/10.1145/1148170.1148219

[6] Alessandro Checco, Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Gianluca Demartini. 2017. Let's agree to disagree: Fixing agreement measures for crowdsourcing. In *Proceedings of the Fifth AAAI Conference on Human Computation and Crowdsourcing*, Steven Dow and Adam Tauman Kalai (Eds.). AAAI Press, 11–20. Retrieved from https://aaai.org/ocs/index.php/HCOMP/HCOMP17/paper/view/15927.

[7] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2020. Offline evaluation without gain. In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR'20).* Association for Computing Machinery, New York, NY, 185–192. DOI: https://doi.org/10.1145/3409256.3409816

[8] Charles L. A. Clarke, Fernando Diaz, and Negar Arabzadeh. 2023. Preference-based offline evaluation. In *Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. Association for Computing Machinery, New York, NY, 1248–1251. DOI: https://doi.org/10.1145/3539597.3572725

[9] Charles L. A. Clarke, Mark D. Smucker, and Alexandra Vtyurina. 2020. Offline evaluation by maximum similarity to an ideal ranking. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management.* Association for Computing Machinery, New York, NY, 225–234. DOI: https://doi.org/10.1145/3340531.3411915

[10] Charles L. A. Clarke, Alexandra Vtyurina, and Mark D. Smucker. 2021. Assessing top-k preferences. *ACM Trans. Inf. Syst.* 39, 3 (2021). DOI: https://doi.org/10.1145/3451161

[11] Jacob Cohen. 1977. *Statistical Power Analysis for the Behavioral Sciences.* Elsevier, Cambridge, MA. DOI: https://doi.org/10.1016/C2013-0-10517-X

[12] Gordon V. Cormack and Thomas R. Lynam. 2007. Power and bias of subset pooling strategies. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07).* Association for Computing Machinery, New York, NY, 837–838. DOI: https://doi.org/10.1145/1277741.1277934

[13] Kalyanmoy Deb, Samir Agrawal, Amrit Pratap, and T. Meyarivan. 2002. A fast and elitist multiobjective genetic algorithm: NSGA-II. *IEEE Trans. Evolut. Computat.* 6, 2 (2002), 182–197. DOI: https://doi.org/10.1109/4235.996017

[14] Fernando Diaz and Andres Ferraro. 2022. Offline retrieval evaluation without evaluation metrics. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval.* Association for Computing Machinery, New York, NY, 599–609. DOI: https://doi.org/10.1145/3477495.3532033

[15] Djellel Difallah, Elena Filatova, and Panos Ipeirotis. 2018. Demographics and dynamics of mechanical turk workers. In *Proceedings of the 11th ACM International Conference on Web Search and Data Mining (WSDM'18).* Association for Computing Machinery, New York, NY, 135–143. DOI: https://doi.org/10.1145/3159652.3159661

[16] Djellel Eddine Difallah, Michele Catasta, Gianluca Demartini, Panagiotis G. Ipeirotis, and Philippe Cudré-Mauroux. 2015. The dynamics of micro-task crowdsourcing: The case of Amazon MTurk. In *Proceedings of the 24th International Conference on World Wide Web (WWW'15).* International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 238–247. DOI: https://doi.org/10.1145/2736277.2741685

[17] Carsten Eickhoff and Arjen P. De Vries. 2013. Increasing cheat robustness of crowdsourcing tasks. *Inf. Retr.* 16, 2 (01 Apr. 2013), 121–137. DOI: https://doi.org/10.1007/s10791-011-9181-9

[18] Paul D. Ellis. 2010. *The Essential Guide to Effect Sizes: Statistical Power, Meta-analysis, and the Interpretation of Research Results.* Cambridge University Press, Cambridge, UK.

[19] Nicola Ferro, Yubin Kim, and Mark Sanderson. 2019. Using collection shards to study retrieval performance effect sizes. *ACM Trans. Inf. Syst.* 37, 3 (Mar. 2019). DOI: https://doi.org/10.1145/3310364

[20] Nicola Ferro and Mark Sanderson. 2022. How do you test a test? A multifaceted examination of significance tests. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining (WSDM'22).* Association for Computing Machinery, New York, NY, 280–288. DOI: https://doi.org/10.1145/3488560.3498406

[21] Nicola Ferro and Gianmaria Silvello. 2016. A general linear mixed models approach to study system component effects. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'16).* ACM, New York, NY, 25–34. DOI: https://doi.org/10.1145/2911451.2911530

[22] Nicola Ferro and Gianmaria Silvello. 2018. Toward an anatomy of IR system component performances. *J. Assoc. Inf. Sci. Technol.* 69, 2 (2018), 187–200. DOI: https://doi.org/10.1002/asi.23910

[23] Andy Field. 2013. *Discovering Statistics Using IBM SPSS Statistics: And Sex and Drugs and Rock 'n' Roll.* Sage, USA.

[24] John Guiver, Stefano Mizzaro, and Stephen Robertson. 2009. A few good topics: Experiments in topic set reduction for retrieval evaluation. *ACM Trans. Inf. Syst.* 27, 4 (Nov. 2009). DOI: https://doi.org/10.1145/1629096.1629099

[25] David Hawking, Ellen M. Voorhees, Nick Craswell, and Peter Bailey. 1999. Overview of the TREC-8 web track—Results. In *Proceedings of the Eighth Text REtrieval Conference (NIST Special Publication, Vol. 500-246)*, Ellen M. Voorhees and Donna K. Harman (Eds.). National Institute of Standards and Technology (NIST), 1–18. Retrieved from http://trec.nist.gov/pubs/trec8/papers/web_results.pdf.

[26] Joel L. Horowitz. 2001. Chapter 52—The bootstrap. *Handb. Economet.* 5 (2001), 3159–3228. DOI: https://doi.org/10.1016/S1573-4412(01)05005-X

[27] Jeff Howe. 2006. The rise of crowdsourcing. *Wired Mag.* 14, 6 (2006), 1–4. Retrieved from https://www.wired.com/2006/06/crowds/.

[28] Tae Kyun Kim and Jae Hong Park. 2019. More about the basic assumptions of t-test: Normality and sample size. *Korean J. Anesthes.* 72, 4 (2019), 331–335. DOI: https://doi.org/10.4097/kja.d.18.00292

[29] Aniket Kittur, Ed H. Chi, and Bongwon Suh. 2008. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI'08).* ACM, New York, NY, 453–456. DOI: https://doi.org/10.1145/1357054.1357127

[30] Klaus Krippendorff. 2008. Computing Krippendorff's alpha-reliability. *UPENN Libr.* 1 (2008), 43. Retrieved from https://repository.upenn.edu/asc_papers/43.

[31] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. expert: What can relevance judgment rationales teach us about assessor disagreement? In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18).* Association for Computing Machinery, 805–814. DOI: https://doi.org/10.1145/3209978.3210033

[32] Justin F. Landy, Miaolei Liam Jia, Isabel L. Ding, Domenico Viganola, Warren Tierney, Anna Dreber, Magnus Johannesson, Thomas Pfeiffer, Charles R. Ebersole, Quentin F. Gronau, Alexander Ly, Don van den Bergh, Maarten Marsman, Koen Derks, Eric-Jan Wagenmakers, Andrew Proctor, Daniel M. Bartels, Christopher W. Bauman, William J. Brady, Felix Cheung, Andrei Cimpian, Simone Dohle, M. Brent Donnellan, Adam Hahn, Michael P. Hall, William Jiménez-Leal, David J. Johnson, Richard E. Lucas, Benoît Monin, Andres Montealegre, Elizabeth Mullen, Jun Pang, Jennifer Ray, Diego A. Reinero, Jesse Reynolds, Walter Sowden, Daniel Storage, Runkun Su, Christina M. Tworek, Jay J. Van

Bavel, Daniel Walco, Julian Wills, Xiaobing Xu, Kai Chi Yam, Xiaoyu Yang, William A. Cunningham, Martin Schweinsberg, Molly Urwitz, The Crowdsourcing Hypothesis Tests Collaboration, and Eric L. Uhlmann. 2020. Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* 146, 5 (May 2020), 451–479. DOI: https://doi.org/10.1037/bul0000220

[33] Eddy Maddalena, Stefano Mizzaro, Falk Scholer, and Andrew Turpin. 2017. On crowdsourcing relevance magnitudes for information retrieval evaluation. *ACM Trans. Inf. Syst.* 35, 3 (Jan. 2017). DOI: https://doi.org/10.1145/3002172

[34] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why is that relevant? Collecting annotator rationales for relevance judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 4, 1 (Sep. 2016), 139–148. Retrieved from https://ojs.aaai.org/index.php/HCOMP/article/view/13287.

[35] Stefano Mizzaro and Stephen Robertson. 2007. Hits hits TREC: Exploring IR evaluation results with network analysis. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'07).* Association for Computing Machinery, New York, NY, 479–486. DOI: https://doi.org/10.1145/1277741.1277824

[36] Michael J. Nelson. 1998. Statistical power and effect size in informative retrieval experiments. In *Proceedings of the Annual Conference of CAIS.* CAIS. DOI: https://doi.org/10.29173/cais437

[37] Ralph G. O'Brien and Mary K. Kaiser. 1985. MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychol. Bull.* 97, 2 (1985), 316. DOI: https://doi.org/10.1037/0033-2909.97.2.316

[38] Stephen F. Olejnik and James Algina. 2004. Generalized eta and omega squared statistics: Measures of effect size for some common research designs. *Psychol. Meth.* 8 (01 2004), 434–47. DOI: https://doi.org/10.1037/1082-989X.8.4.434

[39] Gabriele Paolacci, Jesse Chandler, and Panagiotis G. Ipeirotis. 2010. Running experiments on Amazon Mechanical Turk. *Judg. Decis. Mak.* 5, 5 (2010), 411–419. DOI: https://doi.org/10.1017/S1930297500002205

[40] Javier Parapar, David E. Losada, and Álvaro Barreiro. 2021. Testing the tests: Simulation of rankings to compare statistical significance tests in information retrieval evaluation. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing (SAC'21).* Association for Computing Machinery, New York, NY, 655–664. DOI: https://doi.org/10.1145/3412841.3441945

[41] Flávio Ribeiro, Dinei FlorÃłncio, Cha Zhang, and Michael Seltzer. 2011. CROWDMOS: An approach for crowdsourcing mean opinion score studies. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP).* IEEE, New York, NY, 2416–2419. DOI: https://doi.org/10.1109/ICASSP.2011.5946971

[42] Kevin Roitero, Ben Carterette, Rishabh Mehrotra, and Mounia Lalmas. 2020. Leveraging behavioral heterogeneity across markets for cross-market training of recommender systems. In *Proceedings of the Web Conference (WWW'20).* ACM, New York, NY, 694–702. DOI: https://doi.org/10.1145/3366424.3384362

[43] Kevin Roitero, J. Shane Culpepper, Mark Sanderson, Falk Scholer, and Stefano Mizzaro. 2020. Fewer topics? A million topics? Both?! On topics subsets in test collections. *Inf. Retr. J.* 23, 1 (Feb. 2020), 49–85. DOI: https://doi.org/10.1007/s10791-019-09357-w

[44] Kevin Roitero, Eddy Maddalena, Gianluca Demartini, and Stefano Mizzaro. 2018. On fine-grained relevance scales. In *Proceedings of the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval (SIGIR'18).* ACM, New York, NY, 675–684. DOI: https://doi.org/10.1145/3209978.3210052

[45] Kevin Roitero, Eddy Maddalena, and Stefano Mizzaro. 2017. Do easy topics predict effectiveness better than difficult topics? In *Advances in Information Retrieval*, Joemon M. Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait (Eds.). Springer International Publishing, Cham, 605–611. DOI: https://doi.org/10.1007/978-3-319-56608-5_55

[46] Kevin Roitero, Eddy Maddalena, Stefano Mizzaro, and Falk Scholer. 2021. On the effect of relevance scales in crowdsourcing relevance assessments for information retrieval evaluation. *Inf. Process. Manag.* 58, 6 (2021), 102688. DOI: https://doi.org/10.1016/j.ipm.2021.102688

[47] Kevin Roitero, Michael Soprano, Andrea Brunello, and Stefano Mizzaro. 2018. Reproduce and improve: An evolutionary approach to select a few good topics for information retrieval evaluation. *J. Data Inf. Qual.* 10, 3 (Sep. 2018). DOI: https://doi.org/10.1145/3239573

[48] Tetsuya Sakai. 2016. Topic set size design. *Inf. Retr. J.* 19, 3 (01 June 2016), 256–283. DOI: https://doi.org/10.1007/s10791-015-9273-z

[49] Tetsuya Sakai. 2016. Topic set size design and power analysis in practice. In *Proceedings of the ACM International Conference on the Theory of Information Retrieval (ICTIR'16).* Association for Computing Machinery, New York, NY, 9–10. DOI: https://doi.org/10.1145/2970398.2970443

[50] Tetsuya Sakai. 2018. *Sample Sizes, Effect Sizes, and Statistical Power—Laboratory Experiments in Information Retrieval.* Springer Singapore, 147–148. DOI: https://doi.org/10.1007/978-981-13-1199-4_8

[51] Tetsuya Sakai and Noriko Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Inf. Retr.* 11, 5 (01 Oct. 2008), 447–470. DOI: https://doi.org/10.1007/s10791-008-9059-7

[52] Tetsuya Sakai and Teruko Mitamura. 2010. Boiling down information retrieval test collections. In *Adaptivity, Personalization and Fusion of Heterogeneous Information (RIAO'10)*. Le Centre De Hautes Etudes Internationales D'Informatique Documentaire, Paris, 49–56. DOI: https://doi.org/10.5555/1937055.1937066

[53] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance assessments for web search evaluation: Should we randomise or prioritise the pooled documents? *ACM Trans. Inf. Syst.* 40, 4 (Jan. 2022). DOI: https://doi.org/10.1145/3494833

[54] Parnia Samimi and Sri Devi Ravana. 2014. Agreement between crowdsourced workers and expert assessors in making relevance judgment for system based IR evaluation. In *Recent Advances on Soft Computing and Data Mining*, Tutut Herawan, Rozaida Ghazali, and Mustafa Mat Deris (Eds.). Springer International Publishing, Cham, 399–407. DOI: https://doi.org/10.1007/978-3-319-07692-8_38

[55] Mark Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Found. Trends Inf. Retr.* 4, 4 (2010), 247–375. DOI: https://doi.org/10.1561/1500000009

[56] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proceedings of the 16th ACM Conference on Conference on Information and Knowledge Management (CIKM'07)*. Association for Computing Machinery, New York, NY, 623–632. DOI: https://doi.org/10.1145/1321440.1321528

[57] 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)*. Association for Computing Machinery, New York, NY, 324–330. DOI: https://doi.org/10.1145/564376.564433

[58] 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)*. Association for Computing Machinery, New York, NY, 565–574. DOI: https://doi.org/10.1145/2766462.2767760

[59] Julián Urbano, Matteo Corsi, and Alan Hanjalic. 2021. How do metric score distributions affect the type I error rate of statistical significance tests in information retrieval? In *Proceedings of the ACM SIGIR International Conference on Theory of Information Retrieval (ICTIR'21)*. Association for Computing Machinery, New York, NY, 245–250. DOI: https://doi.org/10.1145/3471158.3472242

[60] William Webber, Alistair Moffat, and Justin Zobel. 2008. Statistical power in retrieval experimentation. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management (CIKM'08)*. Association for Computing Machinery, New York, NY, 571–580. DOI: https://doi.org/10.1145/1458082.1458158

[61] Xinyi Yan, Chengxi Luo, Charles L. A. Clarke, Nick Craswell, Ellen M. Voorhees, and Pablo Castells. 2022. Human preferences as dueling bandits. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'22)*. Association for Computing Machinery, New York, NY, 567–577. DOI: https://doi.org/10.1145/3477495.3531991

[62] Ziying Yang, Alistair Moffat, and Andrew Turpin. 2018. Pairwise crowd judgments: Preference, absolute, and ratio. In *Proceedings of the 23rd Australasian Document Computing Symposium*. ACM, New York, NY, 1–8. DOI: https://doi.org/10.1145/3291992.3291995

[63] Fabio Zampieri, Kevin Roitero, J. Shane Culpepper, Oren Kurland, and Stefano Mizzaro. 2019. On topic difficulty in IR evaluation: The effect of systems, corpora, and system components. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'19)*. ACM, New York, NY, 909–912. DOI: https://doi.org/10.1145/3331184.3331279

[64] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'98)*. ACM, New York, NY, 307–314. DOI: https://doi.org/10.1145/290941.291014

Received 13 September 2022; revised 8 March 2023; accepted 3 May 2023