Further Insights on Drawing Sound Conclusions from Noisy Judgments

DAVID GOLDBERG, eBay
ANDREW TROTMAN, University of Otago
XIAO WANG, eBay
WEI MIN, CreditX
ZONGRU WAN, Evolution Labs

The effectiveness of a search engine is typically evaluated using hand-labeled datasets, where the labels indicate the relevance of documents to queries. Often the number of labels needed is too large to be created by the best annotators, and so less expensive labels (e.g., from crowdsourcing) are used. This introduces errors in the labels, and thus errors in standard effectiveness metrics (such as P@k and DCG). These errors must be taken into consideration when using the metrics. Previous work has approached assessor error by taking aggregates over multiple inexpensive assessors. We take a different approach and introduce equations and algorithms that can adjust the metrics to the values they would have had if there were no annotation errors.

This is especially important when two search engines are compared on their metrics. We give examples where one engine appeared to be statistically significantly better than the other, but the effect disappeared after the metrics were corrected for annotation error. In other words, the evidence supporting a statistical difference was illusory and caused by a failure to account for annotation error.

CCS Concepts: • **Information systems → Retrieval effectiveness**;

Additional Key Words and Phrases: Precision, statistical significance, standard error

ACM Reference format:
David Goldberg, Andrew Trotman, Xiao Wang, Wei Min, and Zongru Wan. 2018. Further Insights on Drawing Sound Conclusions from Noisy Judgments. *ACM Trans. Inf. Syst.* 36, 4, Article 36 (April 2018), 31 pages.
https://doi.org/10.1145/3186195

1 INTRODUCTION

The standard protocol for determining whether one ranking function is superior to another has been developed over a long period of time. In early experiments very little data was shared, making direct comparison difficult. The early publicly available datasets such as CACM and CISI made direct comparison possible. However, it was not until TREC that direct system-to-system comparison was standardized. Each year, TREC publishes document collections and manual assessments of the relevance of documents to queries.

Authors' addresses: D. Goldberg and X. Wang, eBay, 2025 Hamilton Avenue, San Jose, California 95125, USA; emails: {dgoldberg, xwang2}@ebay.com; A. Trotman, Department of Computer Science, University of Otago, P.O. Box 56, Dunedin, New Zealand; email: andrew@cs.otago.ac.nz; W. Min, CreditX, Building G, Bund Soho, Zhongshan East 2nd Rd, Huangpu District, Shanghai, China; email: vera_jack312@126.com; Z. Wan, Evolution Labs, #104 Building 3, Longhua Road 2577, Shanghai, China; email: zwan@evolutionlabs.com.cn.

https://doi.org/10.1145/3186195

These manual assessments are treated as the gold standard ground truth, and the effectiveness of query results is measured with respect to them. Effectiveness metrics such as MAP are used for reporting and statistical tests are performed to determine whether or not differences between runs are significant.

This process relies on the gold standard being accurate. There has been considerable work examining this reliance, and not only in the context of TREC.

In the context of crowdsourcing, many have looked at aggregation approaches to increase the quality of the assessments and even at how many assessors are needed based on the difficulty of the task. But, as Abraham et al. [1] aptly state, measuring the quality of a worker based on conformance to a majority vote which in turn is taken by high-quality workers can lead to "self-fulfilling loops" in which workers who are purportedly skilled all provide the same wrong answer.

If the ground truth is inaccurate (for whatever reason), then the MAP scores are inaccurate, the statistical tests are inaccurate, and the outcome of an experiment comparing two ranking functions might produce an invalid result. We ask:

Is it possible to perform sound statistical significance tests when the assessments contain errors?

We show that it is. We assume two levels of judging, bronze and gold. *Bronze judges* are inexpensive but inaccurate. *Gold judges* are a scarce resource, but have a high level of accuracy. Bronze judge accuracy is estimated by taking a set of bronze judge assessed query–document pairs and asking the gold judge to carefully reassess the pairs. This provides two accuracy rates for the bronze judges: a rate on relevant documents (relevant according to gold judges) and a rate on irrelevant documents. First we ask:

Given the judge accuracy rates, is it possible to compute a more reliable effectiveness score?

We show that it is, and we derive equations for P@1, P@k, and DCG that directly include these accuracy rates, which we believe are essential to accurately measure effectiveness. The equations we derive can be motivated as follows. In the extreme case when ground truth is created by judges making random assessments, then no conclusions can be drawn about the experiment. The opposite extreme occurs when the judges are perfect and conventional statistical tests are correct. Our equations quantify how to interpolate between these two cases—in other words, when judges are not random but do make some errors. The main application of our equations is to statistical significance so we ask:

Is it possible to compute a more reliable standard error?

We show that it is. Then, given the effectiveness and the standard error, we show how to compute a p-value using Welch's version of Student's t-test.

We apply our work to eBay, a large-scale commercial eCommerce search engine, and show that an incorrect conclusion would have been drawn if the accuracy rate were not considered, thus showing that our test is harder to pass than tests not incorporating accuracy rates. We further apply our techniques to a TREC collection that mirrors our environment.

2 RELATED WORK

In this section, we phrase our problem in the context of prior work related to our own.[1] Zobel [33] rigorously demonstrated the robustness of pooling, and we do not question this. Our investigation centers on robust evaluation in an environment of errorful assessments, which others have already demonstrated exists.

Bailey et al. [3] examined three quality levels of judge: gold, silver, and bronze. They found low levels of agreement with the silver and bronze judges and consequently small but consistent

[1] This article is an extended version of our previous conference article [8]. It has been extended with the addition of sections on bias, the boundary case, false negatives, relative rank orderings, assumptions, and power analysis, as well as edits to other sections.

variation in system scores and rankings. Most interestingly, Bailey et al. used bronze judges for 50 topics but the gold judges for only 33 topics, a decision that may have been motivated by cost (highly trained judges cost more than TREC participants). This is exactly our motivation. We cannot afford to train and use high-quality judges for all evaluations. We wish to assess as many queries as possible, so we use bronze judges to assess document–query pairs, then use gold judges to assess the bronze judges. Unlike Bailey et al., we do not have silver judges as it is not obvious how such a judge would fit within our scenario. They were evaluating three separate quality of judge, whereas we evaluate bronze judges using judges we call "gold."

In a cooperative assessment environment such as ours and at TREC, these differences between assessors are not caused by deliberate attack, laziness, or ignorance. Assessors have the time to, and do, read detailed assessment instructions and refer to them throughout the assessment process. However, they continue to disagree. Voorhees [28] identifies some TREC topics where there is no judge agreement on a single document. This disagreement might be genuine disagreement or it might be error. Either way, if it is small then it might safely be ignored; if not, then the assessment process should be carefully examined as it might be systemic in the process, not the humans.

In our case of product search at eBay, we believe that disagreement is less common because product search is akin to entity search, and boundary cases requiring the judge to read the document in detail and make an informed decision are less common than in web (or news) search. Errors, on the other hand, are caused by human factors—humans are fallible no matter how much care is put into the experimental design and how straightforward the task.

Carterette and Soboroff [4] simulate the effect of assessor error under different archetypes of artificial assessor personalities ranging from optimistic and pessimistic to lazy and unenthusiastic. They show that these different personalities of assessors do, indeed, have a marked effect on system evaluation. They suggest that "low-cost evaluation methods are sensitive to noise in the relevant documents" and that the pessimist produces the best overall results, even if metric scores are lower than would be achieved using optimists.

Sanderson et al. [23] investigate the cause of judging errors and show that the relevance of a document to the judge is not independent of previously seen documents due to a number of possible factors, including relevance shift, error, or both. Scholer et al. [24] attribute high levels of judgment error in TREC assessments to many factors, including how far through the assessment process a given judge is (i.e., relevance shift, tiredness, or other factors).

Many have turned to crowdsourcing in an effort to reduce the high cost of generating assessments. Our work applies to both crowdsourcers and more dedicated assessors. Alonso et al. [2] suggested using either a weighted sum or a voting scheme to reduce errors. Snow et al. [26] estimate judge accuracy using examples labeled by an expert judge. They observe that, with enough judges, the accuracy of each judge can be inferred from the others and that a weighted average of all judges could be used as an aggregated assessment (and they also identify an error in their gold assessments). Abraham et al. [1] examine stopping criteria based on the bias seen in the assessment samples and show that when the bias is high enough, the results are often reliable—but warn against assuming the majority vote of crowd workers is the correct answer. Much of the work on using multiple crowd workers assumes a large number of workers, which may or may not be feasible in a practical or cost-constrained production environment.

Vuurens et al. [29] examine spam in crowdsourced assessments. Tang and Lease [27] combine expert assessments with crowdsourced assessments and show that adding small amounts of expert assessment approaches the accuracy of assessments from experts only. Kazai and Zitouni [14] state that obtaining high-quality output from crowd workers remains a key challenge and present a method of identifying bad workers by their behavioral deviation from other workers and gold workers. Dekel and Shamir [7] tackle this problem using an SVM trained on all the assessments as a classifier of bad assessors, which they can do without multiple assessments per example.

Kazai et al. [12] examined judgments used at a large-scale commercial web search engine and identified three types of errors in those judgments: intent mismatch (ambiguity), content quality (documents from unauthoritative sources considered relevant), and judge bias (for example, being too lenient). They went on to point out that disagreement can reflect diversity, and this is therefore not error. Worse, the gold standard may be biased toward one particular interpretation of the query. They give an example of the query "gps" matching shopping sites, gps.gov, and a Wikipedia page. In a shopping vertical such as eBay, the diversity in a query such as "gps" is substantially smaller than on the web; nonetheless, it is important to investigate the cause of disagreement between judges before assuming disagreement is error. Improvements to assessing guidelines, among other changes, can result in a reduction of disagreement. We strongly advocate investigating the cause of disagreement and the use of metrics such as Cohen's kappa or Fleiss's kappa to measure disagreement before assuming disagreement is error. In Section 9, we discuss how to manage non-error disagreement (i.e., diversity).

Kazai et al. suggest that pages from sources such as Wikipedia are harder to assess than some other sources, a problem we simply do not face in a vertical. They also test assessments against click logs; at eBay we use Cranfield evaluation using canned documents, queries, and assessments (which takes seconds) to determine whether an A/B experiment (which takes days) is likely to have a positive outcome or not, and then we run A/B tests. A/B tests are impractical as a method for making incremental improvements to a ranking function. They are, however, essential in measuring whether a change has resulted in an improved user experience.

Dawid and Skene [6] compute the error of a clinician taking discrete observations of a patient. When the answer is known (e.g., diagnosis of a break before an x-ray), the error is easily computed. If the gold standard is not known, then they use the Expectation-Maximization (EM) algorithm to compute the error in each cell of a confusion matrix. Passonneau and Carpenter [19] extend this algorithm by adding a Dirichlet prior.

Zhao et al. [32] were the first to explore both false-positive (type I) errors and false-negative (type II) errors. Their domain is multi-faceted as there can be more than one truth (e.g., name an actor who played Dr. Who). They observe that voting is unreliable as the majority can be wrong, and some sources only include verifiable data and prefer absence of data to incorrect data. We examine type I and type II errors in a single faceted domain for P@1 and P@k and in a multifaceted domain for DCG.

Assessment disagreements between judges might be caused by a number of factors including different relevance drift, ambiguity, or error. It is important to establish the cause of this difference and to make the assessment process more robust where possible. We work with data from an eCommerce platform where (based on anecdotal experience) we believe that disagreement is usually error, and so we are faced with the problem of determining whether one ranking function is superior to another in the light of noisy judgments.

Comparing ranking functions in the light of noise is not new, but previous work considered only noise created by picking a different set of queries or a different set of documents.

Moffat and Zobel [18] and Webber et al. [31] introduce Rank-Biased Precision (RBP) and Rank-Biased Overlap (RBO), respectively, both of which compute an effectiveness score along with a residual. In other words, RBP and RBO give an upper bound and a lower bound on the effectiveness score at any point in the results list. Their uncertainty is uncertainty in the remainder of the results list (assuming no judging errors).

Li et al. [16] use detection theory [17] to study errors of crowdsourced judges. This supposes a very specific model for judge errors using *discrimination* and *bias*. Our work is more general, and we make no assumptions about how judge error arises.

Our contribution is to analyze the effect of judge error and the additional noise it introduces (since judge error is only imperfectly known). We show that ignoring judge error affects conclusions about actual ranking functions on actual datasets.

Noise is usually addressed through the use of a significance test such as the *t*-test or Wilcoxon signed rank test. Smucker et al. [25] examine the use of different significance tests for use with ranking functions and conclude that the *t*-test is as good as bootstrap (and randomization) and better than Wilcoxon's signed rank test. Cormack and Lynam [5] propose the use of the bootstrap for computing the standard error. We build on this work by showing how to apply the bootstrap to compute the standard error in the light of errorful assessments and how to perform Welch's unequal variances *t*-test on the effectiveness score and the standard error.

We assume the judges are not adversarial and are working to build a valid set of relevance assessments. Disagreement, nonetheless, still exists. Voorhees [28] examined the effect of disagreement between multiple judges on the TREC 4 and TREC 6 collections. She shows a generally stable result but notes that differences in MAP as high as 10% can see changes in relative rank order depending on the assessment set. She goes on to examine three possible reasons for general stability: reported results are averages, disagreement happens on boundary documents, and relative position of changes in the results list.

We are particularly interested in product search and use this work at eBay. There, we are interested in both average system effectiveness and improving the effectiveness of smaller sets of difficult queries. Product search is more akin to entity search, where an entity either is or is not correct with respect to the query, and so boundary documents are less common. We are averaging over a large number of queries, but at a low depth (P@3, the number of results on a typical smartphone screen), and so our metric is less stable than MAP (small changes can have a large effect).

There is also prior work on evaluating the accuracy of assessments. Passonneau and Carpenter [19] observe that just because two judges agree, it does not make the agreement accurate. If one judge is usually wrong and the second is sometimes wrong, then no claim can be made about an individual assessment without a confidence interval, so they compute confidence intervals for each assessment. Joglekar et al. [11] compute both the error rate of crowdsource workers and also confidence intervals on their work. From this, it is possible to estimate the accuracy of any answer set and to eliminate bad workers. None of this previous work showed how assessment error affects decisions about the relative effectiveness of two ranking functions.

We give an explicit formula for the effect of assessment error on effectiveness along with the standard error of effectiveness. Together, they give a range of possible effectiveness scores.

Sakai has recently examined topic set size [22] selection and power analysis for IR experiments [21]. We include a section on power analysis in which we show how many assessments are required to achieve a given level of significance.

3 BINARY RELEVANCE ASSESSMENTS

3.1 Precision at 1

This section examines the scenario in which the search engine is given a set of *n* queries and produces a single result for each query. This result is either relevant or nonrelevant. The scenario is not atypical; it is precision at 1 document (P@1), one of the many measures used to indicate the effectiveness of a commercial search engine.

More formally, the probability that this search engine will produce a relevant result is estimated by the average $\bar{j}$,

$$\bar{j} = \sum_{s=1}^{n} \frac{j_s}{n}, \tag{1}$$

Table 1. Notational Summary

m_J	the probability that P@1 is 1, according to the bronze judges
m_G	the probability that P@1 is 1, according to the gold judges
m_{J_R}	the probability that a bronze judge is correct (agrees with the gold judge) when judging a relevant item
m_{J_N}	the probability that a bronze judge is correct (agrees with the gold judge) when judging a nonrelevant item
$\bar{j}$	an estimate of m_J: the fraction of relevant items from a sample of size n, as observed by the bronze judges
$\widehat{m}_{J_R}$	an estimate of m_{J_R}: determined by picking n_R relevant items and having gold judges review the bronze judges' ratings
$\widehat{m}_{J_N}$	an estimate of m_{J_N}: determined by picking n_N nonrelevant items and having gold judges review the bronze judges' ratings

Table 2. Confusion Matrix Showing Possible Outcomes

			Bronze	
			Relevant	Nonrelevant
			m_J	$1 - m_J$
Gold	Relevant	m_G	m_{J_R}	$1 - m_{J_R}$
	Nonrelevant	$1 - m_G$	$1 - m_{J_N}$	m_{J_N}

where n is the number of queries used in the estimate, and j_s is the relevance (1 for relevant, 0 for nonrelevant) of the sth returned document. This is an estimate because n is finite and j_s could differ with a different set of n queries.

The confidence in $\bar{j}$ is given by the standard error σ_j, estimated as $\widehat{\sigma}_j$, using

$$\widehat{\sigma}_j = \sqrt{\frac{\bar{j}(1 - \bar{j})}{n}}. \tag{2}$$

We use j to emphasize that these are relevance estimates of a *judge* (the bronze judge). Equation (1) and Equation (2) assume that the judge's assessments are accurate, which we know they are not.

We assume that there is a well-defined relevance for each document and that there is a gold judge who can reliably determine that relevance value. In Section 9, we discuss this assumption further.

Assume that across the entire assessment set (one assessment per query) the bronze judge gives an assessment that matches that of the gold judge with a probability m_J (we use m as it is a mean, see Section 3.2).

Prior studies [20] have shown that the time to assess a document is different for relevant and nonrelevant documents, so we make no assumption about the probability of a match on documents considered relevant by the gold judge, m_{J_R}, being equal to the probability of a match on documents considered notrelevant m_{J_N}. Table 1 contains a summary of the notation we use.

Table 2 presents the 2 × 2 confusion matrix in gray. The first row is for the case when the document is considered relevant by the gold judge, which occurs with probability m_G. The second row is when the document is not considered relevant by the gold judge, which occurs with probability $1 - m_G$. Similarly, the first column is when the bronze judge assesses as relevant, and the second column is for nonrelevant. The values in the 2 × 2 matrix are *conditional* probabilities; the probability of the bronze judge's decision given the gold judge's decision. For example, the upper left box, m_{J_R}, is the probability that the bronze judge assesses a document relevant, given that the gold judge says it is relevant, and the lower right box, m_{J_N}, is the probability the bronze judge assesses not relevant when the gold judge says it is not relevant.

We are interested in computing m_G, the probability that the one result is considered relevant by the gold judge. Unfortunately, m_G cannot be observed directly. But m_J, the *unconditional* probability that the bronze judge rates the document as relevant, can be observed. The value of m_J is given by

$$m_J = m_G m_{J_R} + (1 - m_G)(1 - m_{J_N}). \tag{3}$$

In words, m_J is the probability that the one returned document is considered relevant by the gold judge and the bronze judge agrees, $m_G m_{J_R}$, plus the probability the document is not relevant according to the gold judge but the bronze judge nonetheless assesses it as relevant, $(1 - m_G)(1 - m_{J_N})$.

From this, m_G, the probability that the gold judge considers the document as relevant, can be computed by first expanding,

$$m_J = m_G m_{J_R} + m_G m_{J_N} - m_G + 1 - m_{J_N}$$

then gathering terms,

$$m_J = m_G(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N})$$

then solving for m_G,

$$m_G = \frac{m_J - 1 + m_{J_N}}{m_{J_R} + m_{J_N} - 1}. \tag{4}$$

The variables m_{J_R} and m_{J_N} are the hidden accuracy rates of the bronze judge but can be estimated from assessments as $\widehat{m}_{J_R}$ and $\widehat{m}_{J_N}$.

An estimate of m_J, $\widehat{m}_J = \bar{j}$, is given by Equation (1). From these, an estimate of m_G, $\widehat{m}_G$, can be computed,

$$\widehat{m}_G = \frac{\bar{j} - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1}. \tag{5}$$

Equation (1) should be compared to Equation (5) as they are both estimates of P@1. In the former, the bronze judge is assumed to be faultless. The latter uses the former along with estimates of the judge's accuracy rates to give a more accurate score. Equation (5) *is* consistent with Equation (1): if $\widehat{m}_{J_R} = \widehat{m}_{J_N} = 1$ then $\widehat{m}_G = \bar{j}$. In other words, if the bronze judge is faultless, then both equations give the same result. At the other extreme, if $\widehat{m}_{J_R} = \widehat{m}_{J_N} = 0.5$, then the bronze judge is performing an action equivalent to a coin toss and we can derive no useful information because the denominator of Equation (5) is 0. $\widehat{m}_G$ and $\bar{j}$ are very different. $\bar{j}$ has a simple formula but is inaccurate, whereas $\widehat{m}_G$ is accurate but has a more complex formula.

3.2 Standard Error

In this section, we compute the variance of $\widehat{m}_G$ (our estimate of P@1). The square root of the variance is the standard error, and statistical significance can be computed from $\widehat{m}_G$ and its standard error. Two different methods for computing the variance are presented: the first is the parametric bootstrap; the second is an explicit equation.

To get the variance of $\widehat{m}_G$ we need to write it as a random variable, which we do by introducing random variables to represent the parts of Equation (5): $\bar{j}$, $\widehat{m}_{J_N}$, and $\widehat{m}_{J_R}$. If J is the random variable which is 1 when a bronze judge assesses a document relevant and 0 otherwise, then $\overline{J} = (J_1 + \cdots + J_n)/n$ is the random variable representing $\bar{j}$. Similarly if J_R is 1 when the bronze judge assesses as relevant a document considered relevant by the gold judge, then $\overline{J}_R = ((J_R)_1 + \cdots + (J_R)_{n_R})/n_R$ represents $\widehat{m}_{J_R}$, similarly for $\widehat{m}_{J_N}$.

ALGORITHM 1: Bootstrap Computation of Standard Error

```
function OneBootstrap($\bar{j}, n$)
    $j^* \leftarrow Binom(\bar{j}, n)/n$
    $j^*_R \leftarrow Binom(\widehat{m}_{J_R}, n_R)/n_R$
    $j^*_N \leftarrow Binom(\widehat{m}_{J_N}, n_N)/n_N$
    $\widehat{m}^*_G \leftarrow (j^* - 1 + j^*_N)/(j^*_R + j^*_N - 1)$
    return $\widehat{m}^*_G$
for $i \leftarrow 1$ to $n.iter$ do
    $samples_i \leftarrow OneBootstrap(\bar{j}, n)$
$var \leftarrow$ Variance($samples$)
return $\sqrt{var}$
```

The random variables J, J_R, and J_N are linked to the variables in the previous section via $\mathbb{E}[J] = \mathbb{E}[\bar{J}] = m_J$, $\mathbb{E}[J_R] = \mathbb{E}[\bar{J}_R] = m_{J_R}$ and $\mathbb{E}[J_N] = \mathbb{E}[\bar{J}_N] = m_{J_N}$, where $\mathbb{E}[J]$ is the expected value of J. This explains the notation m (for mean) introduced in the previous section.

The distribution of the random variable $\bar{J}$ is

$$\bar{J} \approx \frac{\text{Binom}(\bar{j}, n)}{n}.$$

To get the distributions of $\bar{J}_R$ and $\bar{J}_N$, select n_R assessments considered relevant by the gold judge and n_N considered nonrelevant by the gold judge.[2] If the bronze judge rates r of the n_R items relevant, then $\widehat{m}_{J_R} = \frac{r}{n_R}$ is an estimate of J_R, and the distribution of J_R is

$$\bar{J}_R \approx \frac{\text{Binom}(\widehat{m}_{J_R}, n_R)}{n_R}$$

similarly,

$$\bar{J}_N \approx \frac{\text{Binom}(\widehat{m}_{J_N}, n_N)}{n_N}.$$

In all three cases, Binom(θ, n) is the binomial distribution, the number of heads in n coin tosses, where a coin has probability of θ of coming up heads.

The justification is thus: There are only two possible outcomes, relevance or nonrelevance (multi-class assessments are discussed in Section 4.1). We also assume that the judges are independent. In practice, some judges are more accurate than others, and this is discussed in detail in Section 9.

The derivation is thus: If W is a random variable that is 1 with a probability θ and 0 otherwise, then we have a Bernoulli random variable. If we take n observations, then this can be modeled by a binomial distribution Binom(θ, n). The average of the n observations is an estimate of θ, $\widehat{\theta}$, so we model it using the averages, Binom($\widehat{\theta}, n$). But it is the distribution of the average that is of interest,

$$\overline{W} = \frac{W_1 + W_2 + \cdots + W_n,}{n}$$

which is distributed like Binom$(\widehat{\theta}, n)/n$. Applying this to J, J_R, and J_N gives the preceding equations.

[2] This could be a separate set from the n items used to compute $\bar{j}$. That is, it could be a prior process or even the validation phase in a crowdsourcing experiment.

To perform the parametric bootstrap, draw many samples of $\bar{J}$, $\bar{J}_R$, and $\bar{J}_N$ from the binomial distributions, and, for each sample, use Equation (5) to compute an instance of $\widehat{m}_G$, $\widehat{m}^*_G$. The standard deviation of these $\widehat{m}^*_G$ is an estimate of the standard error of $\widehat{m}_G$. The square of the standard deviation is the variance, $\mathbb{V}\,[\widehat{m}_G]$.

The algorithm is presented as Algorithm 1. On Lines 2–5, function OneBootstrap computes a single sample, $\widehat{m}^*_G$, from the binomial distributions. The **for** loop on Line 7 computes a set of samples by calling OneBootstrap on Line 8. Line 9 computes the variance, and Line 10 returns the standard error.

As well as there being a straightforward algorithm for bootstrapping the variance, there is an explicit equation that adds insight. The derivation of the variance, $\mathbb{V}\,[\widehat{m}_G]$, the square of the standard error, is from the standard equation for the variance of a quotient,

$$\mathbb{V}\left[\frac{A}{B}\right] \approx \frac{\mathbb{V}\,[A]}{\mathbb{E}\,[B]^2} + \mathbb{V}\,[B]\,\frac{\mathbb{E}\,[A]^2}{\mathbb{E}\,[B]^4} - 2\frac{\mathbb{E}\,[A]}{\mathbb{E}\,[B]^3}\mathbb{V}\,[A,B]\,,$$

where $\mathbb{V}\,[A,B]$ is the covariance of A and B. To compute $\mathbb{V}\,[\widehat{m}_G]$ set (from Equation (4))

$$A = \bar{J} - 1 + \bar{J}_N$$

and

$$B = \bar{J}_R + \bar{J}_N - 1.$$

We explain below why it is reasonable to assume that $\bar{J}$, $\bar{J}_R$, and $\bar{J}_N$ are uncorrelated (e.g., $\mathbb{V}[\bar{J},\bar{J}_R] = 0$).

Now,

$$\mathbb{V}\,[A,B] = \mathbb{V}\left[\bar{J}_N\right]$$

because

$$\begin{aligned}
\mathbb{V}\,[A,B] &= \mathbb{V}\left[\bar{J} - 1 + \bar{J}_N, \bar{J}_R + \bar{J}_N - 1\right] \\
&= \mathbb{V}\left[\bar{J},\bar{J}_N\right] + \mathbb{V}\left[\bar{J},\bar{J}_N\right] + \mathbb{V}\left[\bar{J}_N,\bar{J}_R\right] + \mathbb{V}\left[\bar{J}_N,\bar{J}_N\right] \\
&= 0 + 0 + 0 + \mathbb{V}\left[\bar{J}_N,\bar{J}_N\right] \\
&= \mathbb{V}\left[\bar{J}_N\right]
\end{aligned}$$

so

$$\mathbb{V}\,[\widehat{m}_G] \approx \frac{\mathbb{V}\left[\bar{J}\right] + \mathbb{V}\left[\bar{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \frac{(\mathbb{V}\left[\bar{J}_R\right] + \mathbb{V}\left[\bar{J}_N\right])(\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} - 2\frac{(\bar{j} - 1 + \widehat{m}_{J_N})\mathbb{V}\left[\bar{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^3}. \quad (6)$$

We set the estimate $\widehat{\sigma}^2_G$ to be the right-hand side of the preceding.

This can be rewritten in a more instructive form via some algebraic manipulations (see the Appendix) as:

$$\widehat{\sigma}^2_G = \frac{\mathbb{V}\left[\bar{J}\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \mathbb{V}\left[\bar{J}_R\right]\frac{(\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} + \mathbb{V}\left[\bar{J}_N\right]\frac{(\bar{j} - \widehat{m}_{J_R})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4}. \quad (7)$$

Equation (7) shows explicitly how the bronze judge accuracy affects the standard error of the estimate of precision at 1, $\widehat{m}_G$. If the assessments are random, then $\widehat{m}_{J_R} \approx \widehat{m}_{J_N} \approx 0.5$ and so $\widehat{\sigma}_G \approx \infty$, as expected.

Prior models assume there is no judge error and consequently $\widehat{\sigma}_G = \mathbb{V}[\bar{J}]$. The rewritten form (Equation (7)) shows that the standard error is in fact larger. First, $\mathbb{V}[\bar{J}]$ is increased by a factor of $(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^{-2}$, and, second, uncertainty in $\widehat{m}_{J_R}$, and $\widehat{m}_{J_N}$ (as measured by $\mathbb{V}[\bar{J}_R]$ and $\mathbb{V}[\bar{J}_N]$) increases $\widehat{\sigma}^2_G$ even further.

In a working environment, it is necessary to compute $\mathbb{V}[\bar{J}]$, $\mathbb{V}[\bar{J}_R]$, and $\mathbb{V}[\bar{J}_N]$. Using variance of $\text{Binom}(\theta, n)/n = \theta(1-\theta)/n$ the variances can be estimated as

$$\begin{aligned}\mathbb{V}\left[\bar{J}\right] &\approx \frac{\bar{j}(1-\bar{j})}{n}\\ \mathbb{V}\left[\bar{J}_R\right] &\approx \frac{\widehat{m}_{J_R}(1-\widehat{m}_{J_R})}{n_R}\\ \mathbb{V}\left[\bar{J}_N\right] &\approx \frac{\widehat{m}_{J_N}(1-\widehat{m}_{J_N})}{n_N}.\end{aligned} \tag{8}$$

The derivation of Equation (7) assumes that J, J_R, and J_N are uncorrelated. They are, in fact, independent, which is even stronger than being uncorrelated. They are independent because knowing a specific value of (say) J_R gives no additional information about the other variables J, and J_N. That is because J is modeled as the bronze judge's assessment of a random document. Knowing that assessment gives no additional information about what happens when a document deemed relevant by the gold judge is randomly selected and rated by a bronze judge (resulting in J_R).

In this section, we derived an equation for the standard error of the estimate $\widehat{m}_G$. Unlike others, we work with the error of $\widehat{m}_G$ (which is an average) rather than trying to determine the error in each individual assessment. This choice was made because inferences about the effectiveness of a ranking algorithm require an estimate of $\widehat{m}_G$, not the error in each assessment.

3.3 P-Value

In an information retrieval ranking experiment, the experimenter is typically trying to determine whether ranking function a outperforms ranking function b. The p-value approach is to assume the ranking functions return equally relevant documents and compute the probability that the observed difference in P@1 under this hypothesis is as large as it is. This section outlines how to compute the p-value using Welch's unequal variances t-test (a variant of Student's t-test).

Equation (5) is the equation for $\widehat{m}_G$, the corrected P@1 score. Equation (7) is the equation for the standard error, $\widehat{\sigma}_G$. Given two ranking functions, a and b, there are two precisions $\widehat{m}_{G,a}$ and $\widehat{m}_{G,b}$, and two standard errors, $\widehat{\sigma}_{G,a}$ and $\widehat{\sigma}_{G,b}$.

First presume that the algorithms are not different (the null hypothesis). Then compute the probability that $\widehat{m}_{G,a}$ and $\widehat{m}_{G,b}$ are different.

If the bronze judge always agrees with the gold judge (error-free assessments), this is estimated via Welch's t-test,

$$t = \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}, \tag{9}$$

where for ranking function a, $\bar{j}_a$ is the P@1 score from Equation (1), s_a^2 is the variance in the n_a assessments. Likewise, $\bar{j}_b$, s_b^2, and n_b for function b.

To compute the p-value for Welch's test, the number of degrees of freedom, ν, is needed. That is given by

$$\nu = \frac{\left(\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}\right)^2}{\frac{\left(\frac{s_a^2}{n_a}\right)^2}{n_a-1} + \frac{\left(\frac{s_b^2}{n_b}\right)^2}{n_b-1}}.$$

The Welch two-tailed p-value is

$$p\text{-value} = 2\Pr(T_\nu \geq |t|),$$

where T_ν is a t-distribution with ν degrees of freedom, and p-value is the probability that the difference in the means is as large as observed. The smaller the p-value, the more likely we will consider algorithm a to have different relevance from algorithm b.

All the necessary parameters to compute the p-value in the light of uncertain assessments are derived in Section 3.1 and Section 3.2. The denominator of Equation (9) is the square root of the squares of the standard errors, which are given in Equation (7). The numerator is the $\widehat{m}_G$ score from Equation (5); so

$$t = \frac{\widehat{m}_{G,a} - \widehat{m}_{G,b}}{\sqrt{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}}}. \tag{10}$$

In Equation (9), t grows like $\sqrt{n}$, meaning t is more likely to be significant as n gets larger. The same is true of Equation (10) because $\widehat{\sigma}^2_{G,a}$ and $\widehat{\sigma}^2_{G,b}$ are both $O(\mathbb{V}[\bar{J}])$ which is $O(1/n)$, so that t is $O(\sqrt{n})$, just as in Equation (9).

In a typical search engine scenario, ν is very large and so we use a normal distribution, Z, as an approximation of the t-distribution, T_ν:

$$p = 2\Pr\left(Z \geq |t|\right)$$

giving

$$p = 2\Pr\left(Z \geq \frac{|\widehat{m}_{G,a} - \widehat{m}_{G,b}|}{\sqrt{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}}}\right). \tag{11}$$

We earlier remarked that although the standard error of Equation (5) can be computed via bootstrap, Equation (7) adds additional insight. As an example, in Equation (9), s_a is the standard error of the judgments and so has the value $\bar{j}_a(1-\bar{j}_a)$. Comparing to the first Equation of (8), this shows that

$$\mathbb{V}\left[\bar{J}_a\right] = s_a^2/n. \tag{12}$$

Now suppose that the bronze judge accuracy is known exactly. Then Equation (7) states that $\widehat{\sigma}^2_{G,a} = \mathbb{V}[\bar{J}_a]/D^2 = s_a^2/(D^2 n)$ where D is the denominator $\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1$. And, from Equation (5), $\widehat{m}_{G,a} - \widehat{m}_{G,b} = (\bar{j}_a - \bar{j}_b)/D$. So, in this case, Equation (9) and Equation (10) give the same answer.

In other words, if the judge error rate is known exactly, then the precision of the ranking functions is shifted, but the statistical significance of whether two ranking functions differ is unchanged. The increase in the size of the confidence intervals comes from uncertainty in the actual error rate of the judges. This makes intuitive sense: The traditional confidence interval (assuming no judge error) represents the range of results you would get with different queries or documents. If judge error is not known precisely, then the interval has to increase to represent the range of possible judge error rates.

3.4 Precision at k

Our discussion has focused on a single document–query pair which has been averaged over n queries resulting in P@1. An equation to compute standard error and the p-value has been given. This section focuses on a results list, specifically precision at k, P@k.

The P@k score according to the bronze judge, $j@k$, for a single query is the number of relevant documents found in the top k results in the results list divided by k.

$$j@k = \frac{\sum_{t=1}^{k} j^{(t)}}{k}$$

where $j^{(t)}$ is the relevance (1 for relevant, 0 for nonrelevant) of the document at position t in the results list.

This score is then averaged over the set of n queries,

$$\bar{j}@k = \frac{\sum_{s=1}^{n} \frac{\sum_{t=1}^{k} j_s^{(t)}}{k}}{n}, \tag{13}$$

where $j_s^{(t)}$ is the relevance of the document at position t of the results list of query s. We call this *column-wise computation* as we first compute the score for each column (i.e., results list), then average over them. However, it is also possible to compute P@k for a set of queries row-wise. To do so, first compute the mean score for position 1 of each results list, then for position 2, and so on, then sum those and average over the depth of the results list, k,[3]

$$\bar{j}@k = \frac{\sum_{t=1}^{k} \frac{\sum_{s=1}^{n} j_s^{(t)}}{n}}{k}.$$

Up to this point, J has represented the first result in the results list, and the bronze judge has assigned a value of $j = 1$ for relevant and $j = 0$ for nonrelevant. For P@k there are k assessments for a results list of length k, represented $J^{(1)}, J^{(2)}, \dots J^{(k)}$.

Precision at k for a single query can now be defined in terms of those k random variables as

$$\bar{J}@k = \frac{1}{k}\left(J^{(1)} + J^{(2)} + \cdots + J^{(k)}\right). \tag{14}$$

Using row-wise evaluation, it is meaningful to compute the expected precision value at each given position in the results list, $\mathbb{E}[J^{(t)}]$, and to compute the expected value of P@k from those,

$$m_J@k = \mathbb{E}\left[\bar{J}@k\right] = \frac{1}{k}\left(\mathbb{E}\left[J^{(1)}\right] + \mathbb{E}\left[J^{(2)}\right] + \cdots + \mathbb{E}\left[J^{(k)}\right]\right).$$

This $\mathbb{E}[\bar{J}@k]$ is computed naively from the bronze judge's assessments. The values $\mathbb{E}[J^{(t)}] = m_J^{(t)}$ on the right-hand side can be written using Equation (3) in terms of m_G, m_{J_R}, and m_{J_N}. Substituting for $\mathbb{E}[J^{(t)}]$ gives

$$\begin{aligned} m_J@k = \frac{1}{k}\Big(&\left(m_G^{(1)}(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N})\right) + \cdots \\ &+ \left(m_G^{(k)}(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N})\right)\Big). \end{aligned}$$

Note that $m_G^{(t)}$ is position-dependent but m_{J_R} and m_{J_N} are position-independent. That is, the precision of the search engine at each position in the results list is different, but the accuracy of the judge is position independent.

The previous equation can be rewritten as

$$m_J@k = \frac{m_G^{(1)} + \cdots + m_G^{(k)}}{k}(m_{J_R} + m_{J_N} - 1) + (1 - m_{J_N}),$$

and pulling all the m_G's to one side of the equation yields

$$m_G@k = \frac{m_G^{(1)} + \cdots + m_G^{(k)}}{k} = \frac{m_J@k - 1 + m_{J_N}}{m_{J_R} + m_{J_N} - 1}. \tag{15}$$

[3] Or $\frac{\sum_{t=1}^{k}\sum_{s=1}^{n} j_s^{(t)}}{n \times k}$, or $\frac{\text{found relevant}}{\text{found}}$; set-wise precision.

The left-hand side of Equation (15) is precision at k using the gold judge's judgment of relevance while the right-hand side uses the observed P@k using the bronze judge's relevance ratings, $m_J@k$, together with m_{J_N} and m_{J_R}.

3.5 Standard Error and P-Value

The similarity between Equation (15) and Equation (4) is striking and means that the standard error of $m_G@k$ has the same form as $\widehat{\sigma}_G^2$ in Equation (7).[4] Specifically, replace $\bar{j}$ in Equation (7) with $\bar{j}@k$ from Equation (14). And replace $\mathbb{V}[\bar{J}]$ with the variance of $\bar{J}@k = (\bar{J}^{(1)} + \cdots \bar{J}^{(k)})/k$. However, this variance can no longer be computed using the first equation of (8), because $J@k$ is not binary. Instead, the variance must be computed directly as a standard deviation of the observed P@k values of the different queries.

The P@k score for a single query can only take k possible values, and, if the number of queries is small, the average P@k will not be close to a normal distribution. This does not affect any of our equations, but it does mean that it may be inaccurate to define the confidence interval as the mean ± a constant times the standard error. In this case, it is better to use a bootstrap confidence interval.

3.6 Boundary Case

The corrected estimate of P@1 given by Equation (5) has what appears to be a subtle flaw in that it can return a probability less than 0 or greater than 1. In this section, we show how to correct for this. A simple solution would be to simply replace Equation (5) with

$$\widehat{m}_G = \max\left(0, \min\left(1, \frac{\bar{j} - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1}\right)\right). \tag{16}$$

This may appear ad-hoc, but is principled. In Equation (3), m_G ranges from 0 to 1 and so m_J has the range

$$1 - m_{J_N} \leq m_J \leq m_{J_R}. \tag{17}$$

This assumes the common situation where the judges are right more than half the time, so that $m_{J_N} + m_{J_R} \geq 1$. The reason that $\widehat{m}_G$ can be outside the interval [0,1] is that Equation (17) may not hold for the estimated values; in other words

$$1 - \widehat{m}_{J_N} \leq \widehat{m}_J \leq \widehat{m}_{J_R} \tag{18}$$

may not hold. In previous sections, we set estimate $\widehat{m}_J$ to $\bar{j}$ and computed the estimates $\widehat{m}_J$, $\widehat{m}_{J_R}$, and $\widehat{m}_{J_N}$ as simple ratios:

$$\widehat{m}_J = \frac{k}{n} \qquad \widehat{m}_{J_R} = \frac{k_R}{n_R} \qquad \widehat{m}_{J_N} = \frac{k_N}{n_N}.$$

The maximum likelihood principle enables us to combine k, n, k_R, n_R, k_N, n_N with Constraint (17) to get estimates for $\widehat{m}_J$, $\widehat{m}_{J_R}$, and $\widehat{m}_{J_N}$ that satisfy Constraint (18).

With no constraints, maximum likelihood leads to simple ratios. For example in the first equation above, $\widehat{m}_J = k/n$ is the value that maximizes the probability of seeing k relevant items out of n; in other words, $\widehat{m}_J$ is the value of z that maximizes

$$f(z) = \binom{n}{k} z^k (1-z)^{n-k}.$$

[4] P@1 is a special case of P@k, so this result is not surprising.

We ensure that $0 \leq \widehat{m}_G \leq 1$ by picking estimates $\widehat{m}_J$, $\widehat{m}_{J_R}$, $\widehat{m}_{J_N}$ simultaneously so that they maximize

$$\begin{aligned}
&\binom{n}{k}\widehat{m}_J^k(1-\widehat{m}_J)^{n-k}\\
&\binom{n_R}{k_R}\widehat{m}_{J_R}^{k_R}(1-\widehat{m}_{J_R})^{(n_R-k_R)}\\
&\binom{n_N}{k_N}\widehat{m}_{J_N}^{k_N}(1-\widehat{m}_{J_N})^{(n_N-k_N)}\\
\text{subject to}\quad&\\
&\widehat{m}_J+\widehat{m}_{J_N}-1\geq 0\\
&\widehat{m}_{J_R}-\widehat{m}_J\geq 0.
\end{aligned}$$

The maximum can be found using Lagrange Multipliers. Using independence and taking logs, we need to maximize

$$\begin{aligned}
&k\log\widehat{m}_J+(n-k)\log(1-\widehat{m}_J)+k_R\log\widehat{m}_{J_R}+(n_R-k_R)\log(1-\widehat{m}_{J_R})\\
&\quad+k_N\log\widehat{m}_{J_N}+(n_N-k_N)\log(1-\widehat{m}_{J_N})+\lambda_1(\widehat{m}_J+\widehat{m}_{J_N}-1)+\lambda_2(\widehat{m}_{J_R}-\widehat{m}_J).
\end{aligned}$$

There are three possible cases depending on which of λ_1 and λ_2 are 0. Setting partial derivatives of the previous equation equal to 0 gives

$$\begin{array}{lllll}
\widehat{m}_J=\dfrac{k}{n} & \widehat{m}_{J_R}=\dfrac{k_R}{n_R} & \widehat{m}_{J_N}=\dfrac{k_N}{n_N} & \lambda_1=0 & \lambda_2=0\\
\widehat{m}_J=\dfrac{k+k_R}{n+n_R} & \widehat{m}_{J_R}=\widehat{m}_J & \widehat{m}_{J_N}=\dfrac{k_N}{n_N} & \lambda_1=0 & \lambda_2\neq 0\\
\widehat{m}_J=\dfrac{k+n_N-k_N}{n+n_N} & \widehat{m}_{J_R}=\dfrac{k_R}{n_R} & \widehat{m}_{J_N}=1-\widehat{m}_J & \lambda_1\neq 0 & \lambda_2=0.
\end{array}$$

The first set of equations holds when Equation (5) gives a value between 0 and 1. In the second set $\widehat{m}_G = 1$, and in the third $\widehat{m}_G = 0$. This is the justification for Equation (16), but it does so in a consistent way by modifying the estimates for $\widehat{m}_J$, $\widehat{m}_{J_R}$, and $\widehat{m}_{J_N}$.

3.7 Type I and Type II Errors

The main result of this article is that when current methods that ignore judge error conclude that the difference between two ranking functions is statistically significant, that might not actually be the case. This section shows that the t value computed using the traditional method is always larger than the t value computed taking judge error into account. More precisely, there cannot be two ranking functions whose relevance scores appear to be noise using current methods but will be considered significant when taking judge error into account.

Analyzing this in more detail, let t_{old} be the conventional method of computing t, and t_{new} be the method proposed here that takes judge error into account. In a hypothesis test, we pick a type I error p we are willing to tolerate, compute a cutoff α_p, and reject the null if $t > \alpha_p$. Since $t_{\text{new}} \leq t_{\text{old}}$, the old and new methodologies only give a different answer when

$$t_{\text{new}} \leq \alpha_p < t_{\text{old.}}$$

When the null is true, then t_{old} incorrectly rejects the null (type I error) but t_{new} does not. So the type I error is reduced. When the null is false, then t_{old} correctly rejects the null but t_{new} does not, so the type II error is increased.

Using Equations (9) and (10), the claim $t_{\text{new}} \leq t_{\text{old}}$ becomes

$$\frac{\widehat{m}_{G,a} - \widehat{m}_{G,b}}{\sqrt{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}}} \leq \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}. \tag{19}$$

To show this, start by squaring both sides,

$$\frac{(\widehat{m}_{G,a} - \widehat{m}_{G,b})^2}{\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}} \leq \frac{(\bar{j}_a - \bar{j}_b)^2}{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}.$$

Factoring out $\gamma = \widehat{m}_{J_R} + \widehat{m}_{J_N} - 1$ from the left-hand side (see Equation (5)), gives

$$\frac{(\bar{j}_a - \bar{j}_b)^2}{\gamma^2(\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b})} \leq \frac{(\bar{j}_a - \bar{j}_b)^2}{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}.$$

Since the numerators are identical, this is the same as

$$\gamma^2\left(\widehat{\sigma}^2_{G,a} + \widehat{\sigma}^2_{G,b}\right) \geq \frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}. \tag{20}$$

Using the definition of $\widehat{\sigma}_{G,a}$ from Equation (7),

$$\gamma^2\widehat{\sigma}^2_{G,a} = \mathbb{V}\left[\bar{J}_a\right] + \frac{\mathbb{V}\left[\bar{J}_R\right](\bar{j}_a - 1 + \widehat{m}_{J_N})^2}{\gamma^2} + \frac{\mathbb{V}\left[\bar{J}_N\right](\bar{j}_a - \widehat{m}_{J_R})^2}{\gamma^2} \geq \mathbb{V}\left[\bar{J}_a\right].$$

It follows from Equation (12) that $\gamma^2\widehat{\sigma}^2_{G,a} \geq s_a^2/n_a$, and similarly for $\gamma^2\widehat{\sigma}^2_{G,b}$ thus demonstrating the correctness of Equation (20) and hence Equation (19). The adjusted t value will be smaller than the unadjusted t value (and there is less likely to be a significant difference). That is, our approach will detect type I errors but not type II errors because it produces t values smaller than before the adjustment.

3.8 Bias

As an estimate of m_G, Equation (5) is biased. This section quantifies that bias.

The standard (second-order) estimate of a ratio of two correlated variables is given by

$$\mathbb{E}\left[\frac{A}{B}\right] \approx \frac{\mathbb{E}[A]}{\mathbb{E}[B]} + \frac{\mathbb{E}[A]\ \mathbb{V}[B]}{\mathbb{E}[B]^3} - \frac{\mathbb{V}[A,B]}{\mathbb{E}[B]^2}$$

in this case, from Equation (5),

$$A = \bar{J} - 1 + \bar{J}_N$$

and

$$B = \bar{J}_R + \bar{J}_N - 1.$$

Now

$$\mathbb{E}[\widehat{m}_G] = \mathbb{E}\left[\frac{A}{B}\right]$$

and

$$m_G = \frac{\mathbb{E}[A]}{\mathbb{E}[B]}.$$

With the bias being given by the difference between the expected value and the actual value,

$$\mathbb{E}[\widehat{m}_G] - m_G \approx \mathbb{E}\left[\bar{J} - 1 + \bar{J}_N\right]\frac{\mathbb{V}\left[\bar{J}_R\right] + \mathbb{V}\left[\bar{J}_N\right]}{\mathbb{E}\left[\bar{J}_R + \bar{J}_N - 1\right]^3} - \frac{\mathbb{V}\left[\bar{J}_N\right]}{\mathbb{E}\left[\bar{J}_R + \bar{J}_N - 1\right]^2}.$$

From Equation (8), it follows that the bias is $O(1/M)$ where M is the minimum of n_R and n_N. Although it is convenient to know the bias, we do not use it in the remainder of this work.

3.9 Relative Rank Order Change

It is pertinent to ask whether the relative rank order of two ranking functions can change as a consequence of considering the assessor accuracy rate. If each topic is assessed by a single assessor, then Equation (5) is a simple linear scaling of precision scores, in which case it is not possible for the relative rank order to differ between the traditional result and the result presented herein. However, a single query might be assessed by more than one assessor.

As an example scenario, consider two assessors and two runs scored with P@1. These two runs might be from the same search engine at different moments in time, and so the same assessor may not be available for reuse. Because we are considering P@1, there might be no overlap between the two runs.

Equation (5) states the P@1 score according to the gold judge is

$$\widehat{m}_G = \frac{\bar{j} - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1}. \tag{21}$$

If the two assessors score the two runs the same P@1, $\bar{j}$, and the accuracy rate on gold non-relevant documents, $\widehat{m}_{J_N}$, is the same, but the accuracy rate on gold relevant documents, $\widehat{m}_{J_R}$, of one assessor is higher than that of another, then the linear scaling for each run is different. In this case, the run judged using the assessor with the highest accuracy rate will be lower than the run judged with the other assessor. That is, the relative rank order can change.

We note that this scenario only happens if a topic is assessed by more than one assessor and their accuracy rates differ. We also note that there is a knock-on effect on the standard error and the p-value.

4 GRADED RELEVANCE ASSESSMENTS

Section 3 discussed the scenario where the judges are asked to mark relevance on a binary scale. In this section, we extend from binary assessments to multi-level (i.e., graded) relevance assessments. This extension can then in turn be used to further extend our methods to DCG.

4.1 Multi-Class Assessments

For graded assessment, the bronze judge is asked to assign one of k possible labels, or grade names $(A_1, \ldots, A_k)$, to each document–query pair. Each label, A_i, has an (unknown) probability, $m_{G_{A_i}}$ (succinctly, m_{A_i}), and the gold judge assigns that label to the document. Just as m_G was unknown in Table 2, each m_{A_i} is unknown; however, the sum of the probabilities m_{A_i} is known to be 1 (every document must be assigned only one label).

Rather than having two accuracy rates, m_{J_R} and m_{J_N}, we now have a square matrix of accuracy rates (a confusion matrix) which we denote $\mathcal{J}$, where $\mathcal{J}_{ji}$ is the probability that the bronze judge said i when the gold judge said j (the binary case is Table 2).[5]

If $\vec{J}$ is a random vector of length k whose ith component is 1 when the bronze judge assigns the label A_i, then the expected output of a bronze judge is

$$\mathbb{E}\left[\vec{J}_i\right] = \sum_{j=1}^{k} m_{A_j} \mathcal{J}_{ji} = (m_A^T \mathcal{J})_i, \tag{22}$$

[5] Note that i is the column index and j is the row index, which differs from convention.

where $(\vec{m}_A^T \mathcal{J})_i$ is the ith element of the product of the vector[6] $\vec{m}_A^T = (m_{A_1}, \ldots, m_{A_k})$ with the matrix $\mathcal{J}$; while $\vec{J}_i$ is the i-th element of the vector $\vec{J}$. In the two-dimensional case with the label A_1 meaning relevant, $\mathbb{E}[\vec{J}_1] = m_J$, where m_J was defined in Equation (3).

Equation (22) can be rewritten as

$$\mathbb{E}\left[\vec{J}\right]^T = m_A^T \mathcal{J} \tag{23}$$

so

$$m_A^T = \mathbb{E}\left[\vec{J}\right]^T \mathcal{J}^{-1}. \tag{24}$$

As with the binary relevance case, the values of $\mathcal{J}$ are not known and must be estimated. The two-dimensional case of Equation (24) reduces to Equation (4) as follows:

$$\begin{aligned}
\mathcal{J} &= \begin{pmatrix} m_{J_R} & 1 - m_{J_R} \\ 1 - m_{J_N} & m_{J_N} \end{pmatrix} \\
\vec{m}_A &= \begin{pmatrix} m_G \\ 1 - m_G \end{pmatrix} \\
\mathbb{E}\left[\vec{J}\right]^T &= \begin{pmatrix} \mathbb{E}[J] & 1 - \mathbb{E}[J] \end{pmatrix}
\end{aligned}$$

so

$$\begin{aligned}
\mathcal{J}^{-1} &= \frac{1}{m_{J_R} + m_{J_N} - 1} \begin{pmatrix} m_{J_N} & m_{J_R} - 1 \\ m_{J_N} - 1 & m_{J_R} \end{pmatrix} \\
\vec{m}_A^T = \mathbb{E}\left[\vec{J}\right]^T \mathcal{J}^{-1} &= \frac{1}{m_{J_R} + m_{J_N} - 1} \begin{pmatrix} \mathbb{E}[J] - 1 + m_{J_N} \\ -\mathbb{E}[J] + m_{J_R} \end{pmatrix}.
\end{aligned}$$

Note that matrix $\mathcal{J}$ is Table 2, that $\vec{m}_A$ are the row labels in Table 2, and that $(\vec{m}_A)_1$ is Equation (4).

4.2 DCG

The graded assessment analogue of precision at k is Cumulated Gain (CG) [10]. Gain is the quantity of relevant material seen by the user, and its exact definition depends on evaluation context. For example, in XCG [13], gain is defined in terms of a quantized score computed for XML document elements assessed on a two-dimensional relevance scale; we leave for further work the extension to subdocument relevance and concentrate here on graded relevance on a whole-document scale measured as the user reads down a results list.

The cumulated gain of a user after they have read k documents (in order) from the results of a single query is

$$\mathrm{CG} = \sum_{s=1}^{k} j^{(s)}$$

where $j^{(s)}$ is the relevance score (given by a judge) of a document at position s (often in the range [0..1], where 0 is not relevant and 1 is most highly relevant).

Eye-tracking experiments have shown that users typically examine a results list top to bottom [9], so it is more important to put a highly relevant document at position 1 in the results list than at position 2, position 3, and so on. Metrics that model such a user model show this decrease in utility by increasingly penalizing a document's contribution to the metric the further down the results list it appears.

[6] We use column vectors, so $\vec{m}_A$ is a $k \times 1$ (column) vector and the transpose $\vec{m}_A^T$ is a row vector.

Position	Query 1	Query 2
1	*Not-relevant* (0)	*Partly-relevant* (0.5)
2	*Highly-relevant* (1)	*Not-relevant* (0)

Fig. 1. Two results lists (relevance grades in braces).

In Discounted Cumulated Gain (DCG), this penalization is done by dividing the score of the document, $j^{(s)}$, by the log of its position in the results list [30]. There are many subtly different interpretations of this (using different bases to the log, and so on). The interpretation we use is

$$\mathrm{DCG} = \sum_{s=1}^{k} \frac{j^{(s)}}{\log_2(s+1)}$$

for the results of a single query.

In the previous section, judges assigned labels $A_1, \ldots A_k$ to queries. For DCG, the labels must be mapped to numerical values, so we introduce the value vector $\vec{v}$ where the label A_j is assigned the numerical value v_j. If $\vec{G}$ is a random vector with a 1 in the tth coordinate when the gold judge assigns the label A_t, then the dot product $\vec{G} \cdot \vec{v}$ is the random variable representing the gold judge's value of the document. The average judgment $\mathbb{E}\,[G]$ is likely to vary with the rank of the documents, so we introduce $G^{(s)}$ for gold judge's judgment of documents at the sth rank, and their value is $\vec{G}^{(s)} \cdot \vec{v}$.

With this notation, the expected DCG according to the gold judge is

$$\mathbb{E}\,[\mathrm{DCG}_G] = \sum_{s=1}^{k} \frac{\mathbb{E}\left[\vec{G}^{(s)}\right] \cdot v}{\log_2(s+1)} = \sum_{s=1}^{k} \frac{\vec{m}_A^{(s)} \cdot v}{\log_2(s+1)}.$$

Repeating Equation (24)

$$m_A^T = \mathbb{E}\left[\vec{J}\right]^T \mathcal{J}^{-1}$$

and substituting gives

$$\mathbb{E}\,[\mathrm{DCG}_G] = \sum_{s=1}^{k} \frac{(\vec{m}_A^{(s)})^T v}{\log_2(s+1)} = \left(\sum_{s=1}^{k} \frac{\mathbb{E}\left[\vec{J}^{(s)}\right]^T}{\log_2(s+1)} \right) \mathcal{J}^{-1} v, \tag{25}$$

which is an estimate of DCG according to the gold judge. Note that $\mathbb{E}\,[\mathrm{DCG}_G]$ is not computed from $\mathbb{E}\,[\mathrm{DCG}_J]$. Instead, it is computed based on how the judges classified documents, $\mathbb{E}[\vec{J}^{(s)}]$, rather than how they were scored, $\mathbb{E}[\vec{J}^{(s)}] \cdot v$.

As an example, suppose we give highly relevant documents a score of 1, not relevant a score of 0, and partly relevant a score of 0.5. Then A^T = (highly relevant, partly relevant, not relevant), and $v^T = (1.0, 0.5, 0.0)$. Suppose the bronze judge gives the ratings shown in Figure 1. We estimate $\mathbb{E}[\vec{J}^{(1)}]^T$ as the vector $(0, 0.5, 0.5)$ since at rank 1 we did not see a rating of highly relevant, and half the time we see partly relevant and half not relevant. Similarly, $\mathbb{E}[\vec{J}^{(2)}]^T \approx (0.5, 0, 0.5)$. Then,

$$\mathbb{E}\left[\mathrm{DCG}_G\right] = \left(\frac{\mathbb{E}\left[\vec{J}^{(1)}\right]^T}{\log_2(1+1)} + \frac{\mathbb{E}\left[\vec{J}^{(2)}\right]^T}{\log_2(2+1)}\right)\mathcal{J}^{-1}v$$

$$\approx \left(\frac{1}{\log_2 2}(0, 0.5, 0.5) + \frac{1}{\log_2 3}(0.5, 0, 0.5)\right)\mathcal{J}^{-1}\begin{pmatrix}1.0\\0.5\\0.0\end{pmatrix},$$

where $\mathcal{J}$ is the confusion matrix (see, for example, the 2×2 case in Table 2).

The standard error in Equation (25) can be computed using the ordinary nonparametric bootstrap (similarly to Cormack and Lynam [5]). Simply select (query, document) pairs without replacement and use Equation (25) on these bootstrap samples.

5 POWER ANALYSIS

In this section, we demonstrate how to do a power analysis. That is, given the level of statistical significance we desire, we compute the number of assessments needed in the sample as well as the number of relevant and nonrelevant reassessments that must be performed.

Before starting the detailed calculations that take judge accuracy into account, we review power analysis for the case when there is no judge error. With the common notation z_α defined by

$$\Pr(Z > z_\alpha) = \alpha$$

and assuming n is large enough so that t is close to normal, rewrite Equation (9) as

$$z_{p/2} = \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}.$$

In other words, if we require the two-tailed p-value to be less than p, we need to pick n_a and n_b to satisfy the preceding. In the common case when $n_a = n_b = n$, this becomes

$$n = \frac{z_{p/2}^2(s_a^2 + s_b^2)}{(\bar{j}_a - \bar{j}_b)^2}. \tag{26}$$

In the case of judge error, after judging, we have values for $\bar{j}$, $\widehat{m}_{J_R}$, $\widehat{m}_{J_N}$, $\mathbb{V}[\bar{J}]$, $\mathbb{V}[\bar{J}_R]$, $\mathbb{V}[\bar{J}_N]$, n, n_R, and n_N. Consequently, we can compute the p-value and determine if two ranking algorithms are significantly different.

For the power analysis, we use the equations to set sizes n, n_R, and n_N which are under the experimenter's control. In other words, we work backward to determine values necessary to achieve (say) $p < 0.05$ that the two ranking algorithms are different.

To detect if two ranking algorithms are significantly different at the level p, start with Equation (10)

$$\frac{|\widehat{m}_{G,a} - \widehat{m}_{G,b}|}{\sqrt{\widehat{\sigma}_a^2 + \widehat{\sigma}_b^2}} > z_{p/2} \qquad \Pr(Z > z_{p/2}) = p/2.$$

Rewrite this as

$$\widehat{\sigma}_a^2 + \widehat{\sigma}_b^2 < \left(\frac{\widehat{m}_{G,a} - \widehat{m}_{G,b}}{z_{p/2}}\right)^2 = \sigma_0^2. \tag{27}$$

This will be satisfied if

$$\begin{aligned} \widehat{\sigma}_a^2 &\leq \sigma_{a,0}^2 \\ \widehat{\sigma}_b^2 &\leq \sigma_{b,0}^2 \\ \widehat{\sigma}_{a,0}^2 + \sigma_{b,0}^2 &= \sigma_0^2 \end{aligned}$$

where $\sigma_{a,0}^2$ and $\sigma_{b,0}^2$ will be chosen later.

Equation (7) for $\widehat{\sigma}_a^2$ and $\widehat{\sigma}_b^2$ has three terms. It is sufficient to have the ith term bounded by $f_i \sigma_{a,0}^2$, as long as $\sum f_i = 1$. So (working just with $\widehat{\sigma}_a$),

$$\begin{aligned} \mathbb{V}\left[\bar{J}_a\right] \frac{1}{(\widehat{m}_{J_{R,a}} + \widehat{m}_{J_{N,a}} - 1)^2} &\leq f_1 \sigma_{a,0}^2 \\ \mathbb{V}\left[\bar{J}_{R,a}\right] \frac{(\bar{j}_a - 1 + \widehat{m}_{J_{N,a}})^2}{(\widehat{m}_{J_{R,a}} + \widehat{m}_{J_{N,a}} - 1)^4} &\leq f_2 \sigma_{a,0}^2 \\ \mathbb{V}\left[\bar{J}_{N,a}\right] \frac{(\widehat{m}_{J_{R,a}} - \bar{j}_a)^2}{(\widehat{m}_{J_{R,a}} - 1 + \widehat{m}_{J_{N,a}})^4} &\leq f_3 \sigma_{a,0}^2. \end{aligned}$$

Write $m_a = \widehat{m}_{J_{R,a}} + \widehat{m}_{J_{N,a}} - 1$, and $\mathbb{V}[\bar{J}_a] = V_a / n_a$ where V_a is the variance of the judgments (s_a^2 in Section 3.3). Similarly $\mathbb{V}[\bar{J}_{R,a}] = V_{R,a}/n_{R,a}$ and $\mathbb{V}[\bar{J}_{N,a}] = V_{N,a}/n_{N,a}$. Solving for the various n:

$$n_a \geq \frac{V_a}{f_1 m_a^2 \sigma_{a,0}^2} \tag{28}$$

$$n_{R,a} \geq \frac{V_{R,a}(\bar{j}_a - 1 + \widehat{m}_{J_{N,a}})^2}{f_2 m_a^4 \sigma_{a,0}^2} \tag{29}$$

$$n_{N,a} \geq \frac{V_{N,a}(\widehat{m}_{J_{R,a}} - \bar{j}_a)^2}{f_3 m_a^4 \sigma_{a,0}^2}. \tag{30}$$

Then repeat for ranking algorithm b. It is common that $n_a = n_b$ and so we make that assumption and so the bounds of Equation (28) are the same for a and b; in other words,

$$\frac{V_a}{f_1 m_a^2 \sigma_{a,0}^2} = \frac{V_b}{f_1 m_b^2 \sigma_{b,0}^2}.$$

Combining this with $\sigma_{a,0}^2 + \sigma_{b,0}^2 = \sigma_0^2$ gives the values for $\sigma_{a,0}$ and $\sigma_{b,0}$:

$$\sigma_{a,0}^2 = \frac{V_a m_b^2}{(V_a m_b^2 + V_b m_a^2)} \sigma_0^2 \qquad \sigma_{b,0}^2 = \frac{V_b m_a^2}{(V_a m_b^2 + V_b m_a^2)} \sigma_0^2.$$

Substituting this into Equations (28)–(30) gives

$$n_a = n_b \geq \frac{V_a m_b^2 + V_b m_a^2}{m_b^2 f_1 m_a^2 \sigma_0^2} \tag{31}$$

$$n_{R,a} \geq \frac{V_{R,a}(V_a m_b^2 + V_b m_a^2)(\bar{j}_a - 1 + \widehat{m}_{J_{N,a}})^2}{f_2 V_a m_b^2 m_a^4 \sigma_0^2}$$

$$\begin{aligned}
n_{N,a} &\geq \frac{V_{N,a}(V_a m_b^2 + V_b m_a^2)(\widehat{m}_{J_{R,a}} - \bar{j}_a)^2}{f_3 V_a m_b^2 m_a^4 \sigma_0^2}\\
n_{R,b} &\geq \frac{V_{R,b}(V_a m_b^2 + V_b m_a^2)(\bar{j}_b - 1 + \widehat{m}_{J_{N,b}})^2}{f_2 V_b m_a^2 m_b^4 \sigma_0^2}\\
n_{N,b} &\geq \frac{V_{N,b}(V_a m_b^2 + V_b m_a^2)(\widehat{m}_{J_{R,b}} - \bar{j}_b)^2}{f_3 V_b m_a^2 m_b^4 \sigma_0^2}.
\end{aligned}$$

That is, for any two runs and a given p-value, there is a guideline on how many bronze assessments must be checked by gold judges. In practice, it is difficult to get exactly the required number of relevant and nonrelevant assessments required, so the gold judge must continue assessing until *all* the minimums have been reached.

Taking into account Equation (27), both Equation (26) and Equation (31) have in common that n is proportional to $z_{p/2}^2$.

6 SIMULATION

This section uses simulation on a specific example to illustrate that the traditional and corrected equations for P@k and its standard error can give dramatically different results. Our approach is to simulate a judge that makes errors; for a contrasting approach, see Carterette and Soboroff [4], who take assessments and perturb them based on assumed assessor personalities.

Imagine a search engine with a true but hidden P@k of 0.4. This is hidden because it is not practical to assess all possible queries. Instead, we need to rely on imperfect bronze judges who only assess a subset of queries.

Further imagine that the true but hidden judge accuracy rates are $m_{J_R} = 0.9$ and $m_{J_N} = 0.8$. These are hidden because it is not possible to compute the agreement level over all possible assessments; only a sample of the bronze judge's assessments is given to the gold judge for evaluation.

Take a ranking experiment involving $n = 50$ queries where the search engine returns the top $k = 10$ results per query. In order to estimate $m_{J_R} = 0.9$ and $m_{J_N} = 0.8$, the gold judge assesses documents until they have assessed $n_R = 250$ of the documents as relevant and $n_N = 250$ of the documents as nonrelevant. These documents are then given to bronze judges in order to compute the estimates $\widehat{m}_{J_R}$ and $\widehat{m}_{J_N}$.

Knowing the true (hidden) values and the number of samples makes it possible to simulate scores for both P@k and the corrected P@k. The algorithm is presented as Algorithm 2. The first input parameter m_G is a vector, with $m_G[s]$ the average precision at rank s. Line 3 sets up an array g with $g[s,t] = 1$ when the document of rank s for the tth query is relevant. Lines 4–9 compute the relevance that the bronze judge will report, taking into account the judge accuracy probabilities m_{J_R} and m_{J_N}. Lines 10–12 compute the bronze judge's estimate of P@k. Lines 13–14 compute the estimates of m_{J_R} and m_{J_N} that will be obtained by asking the bronze judges to assess n_R and n_N documents, respectively. Line 15 is the corrected value that would be computed using Equation (15).

This simulation was run 10,000 times, and the P@k scores are displayed as a histogram in Figure 2. The red histogram is P@k computed using the traditional (naive) method. The blue histogram is P@k computed according to Equation (15). Even though bronze judge accuracy is high, the naive computation of P@k consistently returns values higher than the true value of 0.4.

The simulation also allows us to compare the naive computation of the standard error with Equation (7). On each simulation, we test if the 95% confidence interval about the estimate for P@k contains the true value of 0.4. Using the naive standard error (the standard deviation of the

ALGORITHM 2: Simulation

```
**function** ONE_SIMULATION($m_G, m_{J_R}, m_{J_N} n_R, n_J, n, k$)
    **for** $s$ **in** $1\!:\!k$ **do**                                   ▷ $g[s,] = 1$ with prob $m_G[s]$
        $g[s,] \leftarrow Bernoulli(m_G[s], n)$
    $j \leftarrow g$                                                  ▷ what bronze report
    **for** $(s, t)$ **in** $(1\!:\!k, 1\!:\!n)$ **do**                          ▷ bronze errors
        **if** $g[s, t] = 1$ **then**
            $j[s, t] \leftarrow 0$ with prob $1 - m_{J_R}$
        **else**
            $j[s, t] \leftarrow 1$ with prob $1 - m_{J_N}$
    **for** $t$ **in** $1\!:\!n$ **do**                                   ▷ judge's P@k for each query
        $\bar{j}[t] \leftarrow average\ j[, t]$
    $j^* \leftarrow$ ave $\bar{j}$                                        ▷ average $\bar{j}[t]$ over all $t$
    $\widehat{m}_{J_R} \leftarrow Binom(m_{J_R}, n_R)/n_R$
    $\widehat{m}_{J_N} \leftarrow Binom(m_{J_N}, n_N)/n_N$
    $g^* \leftarrow (j^* - 1 + \widehat{m}_{J_N})/(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)$
    **return** $j^*, g^*$
```

10 numbers $\bar{j}[t]$ divided by $\sqrt{10}$), the confidence interval includes the true value only 5% of the time. So the confidence interval is much smaller than it should be. Using Equation (7), the interval contained the true value 95% of the time, as it should.

In the simulation, the true value of P@k is known to the simulation but not to the experimenter who is evaluating the search engine. Each run of the simulation returns the P@k that would be observed by the experimenter in an actual experiment. In every run, the naive observed P@k (red) is larger than the true value, often substantially so.

7 USE IN A LIVE ENVIRONMENT

This work is in use at eBay with its large-scale eCommerce search engine. Unlike a web archive, an eCommerce environment such as ours is highly dynamic: due to inventory changes, the document collection is not static; due to changing user needs, the query set is not static; due to changes in the code-base, the ranking function is not static. We wish to know whether, in this environment, there has been a statistically significant change in effectiveness due to ranking function changes.

We took a sample of the eBay query log (13 March 2015 to 20 March 2015) and removed nonsensical queries such as "index.php option com k2" and informational queries such as "how to cancel a bid." From this, we randomly selected 6,787 queries weighted by frequency. These queries included a selection of sort types including Best Match, Total Cost Low to High (including postage), Total Cost High to Low, Time Ending Soonest, and so on. A typical such query is "womens khaki shorts size 4" ordered on price low to high. The length of queries typically ranges from 1 term to 36 terms, with a median of 3.0 and a mean of 2.9.

For each query, we selected the top three results (or fewer if recall was less than three) and sent those query–document pairs for assessment by a professional third party. We provided a detailed set of assessment instructions explaining exactly how to judge, including edge cases. An example edge case *Hot Melt Glue Gun with Glue Sticks* is instructed to be judged not relevant to the query *Hot Glue Gun Sticks* because it includes a gun as well as sticks.

The judges were asked to give a relevance assessment of the item with respect the query. Available options were: relevant, not relevant, will not judge, and unclear intent. Judges could choose

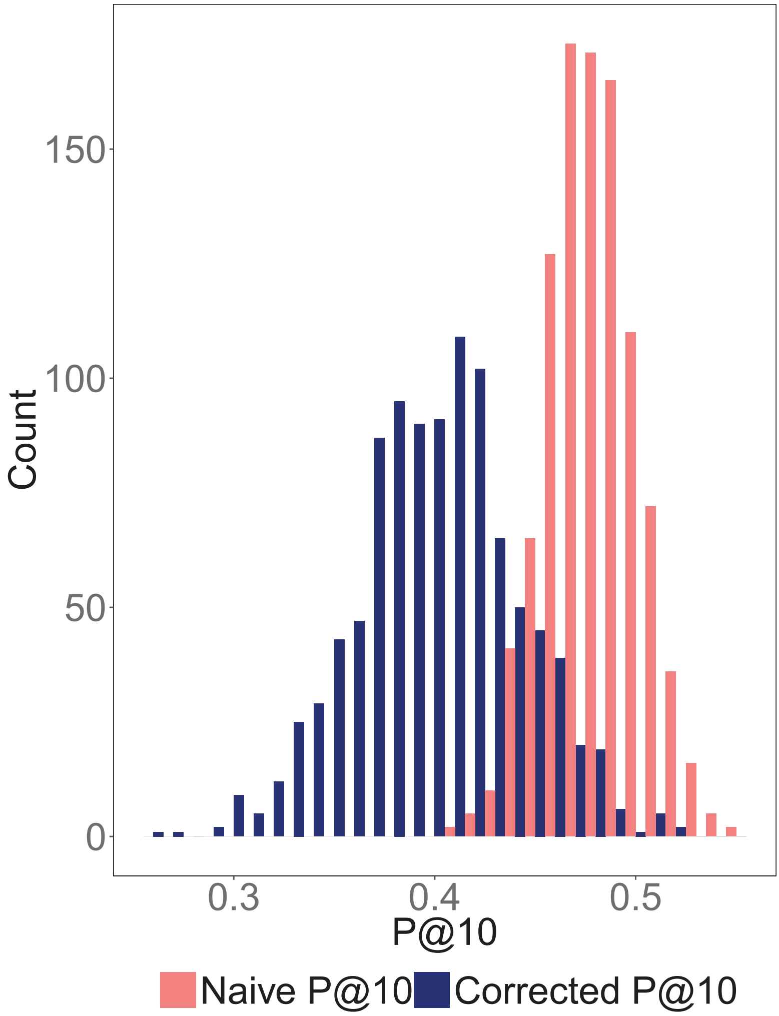

Fig. 2. Histogram of simulated P@10 scores and corrected P@k scores. The naive P@K scores computed by averaging over 50 queries (red) are almost always much higher than the true value of 0.4. Algorithm called as ONE_SIMULATION([0.49, 0.47, . . . 0.31]), 0.9, 0.8, 250, 250, 50, 10).

not to judge a document–query pair and did so for queries such as “t shirt women bikini” and for explicit queries. Judges assigned unclear intent to queries such as “fet irf5307,” for which we note Google returns no result (but is possibly a transistor).

It is pertinent to ask what “relevance” means in the context of product search, and what it does not mean. The assessor is performing a nontrivial matching process requiring detailed instructions, which often requires domain knowledge. For example, the item “Grandt Line O Scale (1:48)” is considered relevant to the query “Grandt O 30” because O is a model railroad abbreviation for 1:48 scale reduction. Such expertise is provided by the third party who employ full-time judges paid by the hour (not by the assessment).

In total, 32 assessors were used, 3 per document–query pair. From this assessment set, we removed all instances of nonassessment (will not assess, unclear intention, and cases where the assessor did not assess), leaving 59,360 assessments. We then computed Fleiss’s kappa. Fleiss’s kappa

Table 3. Statistics of Initial Sampling

Queries	6,787
Assessments	59,360
Assessed by 1 judge (discarded)	251
Assessed by 2 judges	1,330
Both assessors agree	1,034
Both assessors disagree	296
Assessed by 3 judges	18,900
All three agree	13,707
All three disagree	5,193
Fleiss' kappa	0.514

is similar to Cohen's kappa, but is used in the case where there are multiple assessors assigning categorial ratings and not all assessors have necessarily given an assignment to each item. In this case, 3 of the 32 assessors assigned labels to each of the document–query pairs, and so Fleiss's kappa is a more appropriate measure than Cohen's kappa.

For this dataset, there were 251 pairs only assessed by 1 assessor and those were removed from the kappa computation. There were 296 pairs with only 2 assessments and complete disagreement (and 1,034 with agreement) that were included in the computation. In total, we used 14,741 pairs in which there was total agreement, and 5,489 pairs in which there was disagreement. Fleiss's kappa was 0.514, which, according to Landis and Koch [15], represents "Moderate" agreement among the bronze judges. These statistics are restated in Table 3.

We resampled the query log, this time randomly selecting 3,426 queries (from a two-week period in January 2015). These were again sent out for assessment of the top three documents. In total, approximately 30 judges were used. We used the Carterette and Soboroff [4] supermajority vote[7] as the final assessment of the query–document pair (according to the bronze judge) as they suggest this results in a more pessimistic assessment set that is more stable.

Approximately ten days later, the process was repeated (reselection of queries—this time 6,868 from a two-week period in February 2015—and top three results, the same third party might have used different judges), resulting in two datasets,[8] a and b, the characteristics of which are:

$$n_a = 10278 \quad \bar{j}_a = 0.6260 \quad s_a = 0.414$$
$$n_b = 20604 \quad \bar{j}_b = 0.6385 \quad s_b = 0.402$$

$\bar{j}$ is the average P@3 as measured by the judges,[9] and s_a, s_b are the standard deviations of the P@3 scores. The naive estimate of the p-value comes from Equation (9):

$$t = \frac{\bar{j}_a - \bar{j}_b}{\sqrt{\frac{s_a^2}{n_a} + \frac{s_b^2}{n_b}}}$$
$$p\text{-value} = 2(1 - \Pr(T_\nu < |t|)) \approx 0.011,$$

[7] In this case, at least two votes for relevance.

[8] eBay is a large-scale commercial search engine in which there are many A/B experiments concurrently being conducted and in which the ranking function is constantly being changed. As such, it is reasonable to assume change to the ranker over the sample periods—and there was certainly change to the document collection. We do, however, believe that the samples are representative and unbiased.

[9] For web search, these scores are low, but in eCommerce it is common to see queries for items no longer in the inventory.

suggesting a significant (in the statistical sense) change in effectiveness. To verify this, we randomly sampled 143 assessments from the judges and gave those to an in-house expert who acted as the gold judge. Of the 59 the gold judge found relevant, the bronze judge agreed on 43 ($\widehat{m}_{J_R} = 0.729$). Of the 84 the gold judge found nonrelevant, the bronze judge agreed on 67 ($\widehat{m}_{J_N} = 0.798$).

Using Equation (5) to compute corrected P@3,

$$\widehat{m}_{G,a} = \frac{\bar{j}_a - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1} = 0.805$$

$$\widehat{m}_{G,b} = \frac{\bar{j}_b - 1 + \widehat{m}_{J_N}}{\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1} = 0.828.$$

Note that $|\widehat{m}_{G,a} - \widehat{m}_{G,b}| > |\bar{j}_a - \bar{j}_b|$, suggesting that, after correction, the change between a and b might be more significant than previously thought. However, using Equation (7) to update the standard errors (see Section 3.5),

$$\begin{aligned}\widehat{\sigma}^2_{G,a} = &\left(\frac{s_a^2}{n_a}\right)\frac{1}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2}\\ &+ \left(\frac{\widehat{m}_{J_R}(1 - \widehat{m}_{J_R})}{n_R}\right)\frac{(\bar{j}_a - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4}\\ &+ \left(\frac{\widehat{m}_{J_N}(1 - \widehat{m}_{J_N})}{n_N}\right)\frac{(\bar{j}_a - \widehat{m}_{J_R})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4}\\ &\approx 0.0903^2\end{aligned}$$

and likewise, $\widehat{\sigma}_{G,b} \approx 0.0923$.

The corrected estimate of the p-value from Equation (11),

$$p\text{-value} = 0.841,$$

suggests that there was no statistically significant change in effectiveness over the period of experimentation.

In summary, when the accuracy rates were not included in the evaluation, we were led to the incorrect conclusion that a statistically significant change had occurred, but by including the accuracy rates, the reality proves otherwise. We believe this example demonstrates one compelling reason why it is essential to include accuracy rates from now on.

8 USE ON TREC DATA

In 2007, one of the TREC Enterprise Track tasks was finding key documents that might be useful in generating an overview of a given subject. The assessments were community generated by bronze judges. The organizers used gold judges (topic originators and experts in the field) to reassess 33 of the 50 topics [3]. There were 3,004 reassessed query–document pairs. No gold judge inter-annotator agreement was computed. We make the assumption that all assessment differences between the gold judges and the bronze judges is error (an overestimate); however, some is likely to be genuine disagreement, and we leave for further work the reassessment of this collection to differentiate the disagreement rate from error rate.

We wish to know whether there is a statistically significant difference at the 5% level between the top two runs (by P@20).

Table 4 presents the $\bar{j}$@20 numbers obtained by the bronze judges and the accuracy of bronze judges based on reassessment by the gold judge. *If there was no judge error*, then the number of

Table 4. TREC Enterprise Results for the Top Two Runs

TREC run	$\bar{j}$@20	s	n	$n_{J,R}$	$n_{G,R}$	$\widehat{m}_{J_R}$	$n_{J,N}$	$n_{G,N}$	$\widehat{m}_{J_N}$
DocRun02	0.527	0.240	50	17	38	0.447	216	262	0.824
york07ed4	0.513	0.260	50	14	50	0.280	230	285	0.807

The column $\bar{j}$@20 is the naive estimate of P@20, and s is the variance of the n judgments. The other columns quantify the accuracy of the bronze judges. The number of documents deemed relevant by the gold judges is $n_{G,R}$. Of those, $n_{J,R}$ are judged relevant by the bronze judges. Columns $n_{G,N}$ and $n_{J,N}$ are for nonrelevant documents.

judgments needs to determine if the top two runs were different, with $p = 0.05$ given by Equation (26) as[10]

$$n = \frac{z_{p/2}^2(s_a^2 + s_b^2)}{(\bar{j}_a - \bar{j}_b)^2} = \frac{1.96^2(0.240^2 + 0.260^2)}{(0.527 - 0.513)^2} \approx 2453.$$

The very low accuracy rate of relevant assessments, $\widehat{m}_{J_R}$, made by the bronze judges is striking. Being less than 50% in both runs suggests that the estimate $n = 2{,}453$ may be a gross underestimate *when judge error is taken into account.* That is qualitatively confirmed by Equation (7), which shows that the variance of $\bar{j}$@20 is proportional to a power of $(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^{-1}$, which is large when $\widehat{m}_{J_R}$ is small.

In principle, the equations of Section 5 can be used to compute how many judge assessments are needed to draw any sound conclusions. However, the quality of the judge estimates in Table 4 are too low. Performing quality control using Equation (17) (recall the notation: $\bar{j}$@20 is our estimate of m_J), we must have $m_J \leq m_{J_R}$ so we expect

$$\bar{j}_a@20 \lessapprox \widehat{m}_{J_{R,a}} \qquad \bar{j}_b@20 \lessapprox \widehat{m}_{J_{R,b}}.$$

From Table 4

$$0.527 \lessapprox 0.447 \qquad 0.513 \lessapprox 0.280,$$

which is nonsensical. The TREC data as summarized in Table 4 are simply not consistent. If we attempt to use the equations of Section 5, we find from using Equation (16) that $\widehat{m}_{G,a} = \widehat{m}_{G,b} = 1$, and so from Equation (27) $\sigma_0 = 0$, and so from Equation (31) all the estimates for n are infinity.

9 ASSUMPTIONS

The purpose of this article is to study how judge accuracy affects the estimation of effectiveness and the comparison of different search engines. In order to study judge accuracy rigorously, we must make some assumptions. The assumptions we use are:

- There is a well-defined relevance for each document, and there is a gold judge who can reliably determine that relevance value
- The judging process is a black box
- Judgments are independent
- The gold judges review a sample of bronze judgments that are representative of the overall set of bronze judgments

In some industrial applications, these are close to being satisfied. Commercial companies often have very detailed guides defining relevance, so the first assumptions is reasonable. It is common

[10] This is only approximate, because it assumes that when more judgments are done, the values $\bar{j}_a$, $\bar{j}_b$, s_a, and s_b are unchanged.

to outsource human judgment to a third party, in which case the second assumption is reasonable. The fourth assumption becomes important when gold assessments are done once and then used to evaluate future bronze judgments. If the quality of the bronze judges changes over time, the gold judgments may become out of date.

Let us examine each of these assumptions in turn. There is a substantial body of work on user intent and query disambiguation which has clearly shown that different users may express different information needs with the same query, and the same user may even express different intents at different times but with the same query. Nonetheless, there are some important situations (including product search at eBay) where the first assumption is a good approximation. The simplest case is when relevance is defined by a specific user, and the gold judge is the user. In this case, it is obvious that there is a well-defined relevance and that the gold judge (the user) reliably determines it. It is still true that the user's idea of relevance might change. But, as mentioned in Section 2, product search is a form of entity search which is simpler than general search, and we expect such changes over time to be rare. In this simple case, an experimenter wants to know the relevance of an algorithm (or to compare two algorithms). The experimenter hires inexpensive bronze judges to do extensive judgments and personally judges a sample of the bronze judge's work, thereby obtaining $\widehat{m}_{J_N}$ and $\widehat{m}_{J_R}$ of Equation (5), which enables the user to estimate the relevance as if they had done all the evaluations personally.

More generally, the first assumption is reasonable whenever the judges represent a "user," in other words, a single point of view on what relevance means. Industrial search engines like eBay fall into this situation. An executive or committee develops a point of view for documents whose relevance might vary from person to person. We mentioned an example in Section 7. For the query *Hot Glue Gun Sticks*, some users might find *Hot Melt Glue Gun with Glue Sticks* to be relevant, others not. The "user" makes a decision and encodes that in a document guideline which is consulted by the gold judges. Of course, the guidelines cannot cover every ambiguous query. But as disagreements between the gold judges are discovered, the "user" discerns a principle applying to those disagreements and adds it to the guidelines.

You might wonder if this is all sophistry, designed to draw attention away from the very real fact that different users have different ideas about relevance, even in the more constrained arena of entity search. We do not believe so, and illustrate this with a specific example. Are *basketball shoes* relevant to the query *basketball*? Users of eBay will have differing opinions. But the "user" represented by the guidelines does have a point of view, which is that shoes are not relevant. The "user" reasons as follows. A visitor to eBay hoping to find basketball shoes in response to the query *basketball* can easily refine the query to *basketball shoes*. But if shoes were considered relevant, it is much harder for a visitor who wants the actual balls to reformulate the query to remove shoes. When eBay measures relevance, it is measuring relevance for the "user", as represented in the guidelines. Some of the ambiguous relevance decisions will differ from those of a visitor. But, assuming a competent "user," the visitors with differing notions of relevance are still well-served by the site, and it does make sense to measure relevance as determined by the "user", and approximated by gold judges who interpret the "user"'s wishes via the guidelines. In short, visitors to eBay are trained by the "user" to ask for what they want, and in turn the system returns what they ask for—something more possible in an entity search system than an informational system.

That all said, in web search and other domains, ambiguity can be far more prevalent. That is, if there were more than one gold judge, then we do not expect total agreement among them. This can be seen in the Cystic Fibrosis collection, in which experts were asked to debate documents they disagreed on but never came to a conclusion on some. This kind of disagreement is not error, it is valid disagreement caused by (among other things) ambiguity in the interpretation of the query or the intention of the user. Herein, we do not distinguish between valid disagreement and error; we assume all disagreement is error (indeed, there is only one gold judge on the TREC data and so valid disagreement may explain our result there). In a more ambiguous environment than product search, it is essential to distinguish between error and ambiguity. Not doing so results in an overestimate of the error, larger confidence intervals, and a lower likelihood of finding significant difference where one actually exists.

We leave for future work the question of how, exactly, to distinguish between ambiguity and error, along with how to incorporate that into the equations we present. Intuitively and simplistically, by using multiple gold assessors it should be possible to distinguish between genuine error in the bronze judge (where all the gold judges agree that bronze is in error) and differences of opinion (where some gold judges agree with the bronze judges and others do not).

The second assumption (black box) means that there is no metadata associated with a label that lets us determine that one label is more likely to be correct than another. This assumption is needed because we summarize the accuracy of the judgment process in the numbers m_{J_R} and m_{J_N}, which are independent of any document. This might seem unreasonable since some judges are more accurate than others. But numbers like m_{J_R}, which refer to the relevance labels from bronze judges, do not have to literally be from a single judge. The bronze label can come from any judgment process, for example taking the majority vote from three actual bronze judges. Then m_{J_R} and m_{J_N} are the accuracy of the process. In this case, the black box assumption is more reasonable since judgment processes are designed to compensate for judges of lesser quality.

A simple example illustrates this. Suppose each document is evaluated by a single judge, and one judge is less reliable than the others. And suppose we knew which judge evaluated each document. Then our model would not be appropriate because we assume that there is a single fixed probability that the judgment of a relevant document is correct, m_{J_R}. If we know which documents were assessed by the bad judge, there would be two probabilities, not one. This issue goes away if we use majority vote from three assessments. Then all judgments have the same probability of being correct. This is an oversimplified example but illustrates the purpose of processing raw judgments to get a more accurate judgment. Another common process that makes the black box assumption reasonable is monitoring the judge's accuracy. When a judge is determined to have low accuracy, that judge can be removed.

Black box does not mean the assessors have not gone through a rigorous training program in order to be able to make informed decisions. It does not mean that the judges' quality is not monitored. It merely means that correction for poor judges has been done during the judging process and that these adjustments do not have to be taken into account when computing m_{J_R} and $\mathbb{V}[\overline{J}_R]$. We leave for further work the identification of error types in the assessments and the incorporation of that into more sophisticated equations that compensate for judge error.

Our assumption that judgments are independent simply states that the assessment process assumes that each document–query pair is assessed irrespective of other results that might have been seen beforehand. This is a common assumption and is also seen in TREC assessment. It is simply too hard to make a reusable test set in which the relevance of a document is dependent on whether or not the user has seen a prior document. The independence assumption makes it possible to treat the binary assessment as a single binomial process.

10 CONCLUSION

Relevance assessments usually have errors, and it is often not practical to correct all the errors by doing extra assessments or getting more accurate assessors. Our work applies to this common situation. We derive methods that can adjust for assessment errors. We only require that a sample of the assessments be more accurately rejudged.

We show how to adjust P@1, P@k, and DCG, along with their standard errors. From these, we compute a p-value that can be used to test for statistical significance.

A simulated example demonstrated how our adjustments can recover the value the metric would have had if the assessments did not have errors. We applied the adjustments to our search engine and determined that previously believed improvements were illusory and due to bronze judge inaccuracy. Finally, we applied our metrics to TREC data and showed that insufficient data was available to draw conclusions.

We believe that the extra work necessary to evaluate the accuracy of the bronze judges is worthwhile as it allows us to draw sound conclusions from the necessarily noisy assessments.

A DERIVATION OF EQUATION (7)

Start with the right hand side of Equation (6) and collect $\mathbb{V}[\bar{J}_N]$, giving

$$\widehat{\sigma}_G^2 = \frac{\mathbb{V}\left[\bar{J}\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \frac{\mathbb{V}\left[\bar{J}_R\right](\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} + S, \tag{32}$$

where S is

$$\mathbb{V}\left[\bar{J}_N\right]\left(\frac{1}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2} + \frac{(\bar{j} - 1 + \widehat{m}_{J_N})^2}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} - 2\frac{\bar{j} - 1 + \widehat{m}_{J_N}}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^3}\right).$$

Then expand S as

$$S = \frac{\mathbb{V}\left[\bar{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4}\Big((\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2 + (\bar{j} - 1 + \widehat{m}_{J_N})^2 - 2(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)(\bar{j} - 1 + \widehat{m}_{J_N})\Big)$$

or equivalently,

$$S = \frac{\mathbb{V}\left[\bar{J}_N\right]}{(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^4} S_1, \tag{33}$$

where

$$S_1 = (\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)^2 + (\bar{j} - 1 + \widehat{m}_{J_N})^2 - 2(\widehat{m}_{J_R} + \widehat{m}_{J_N} - 1)(\bar{j} - 1 + \widehat{m}_{J_N}).$$

Rewrite S_1 with $z = \widehat{m}_{J_N} - 1$ to get

$$S_1 = (\widehat{m}_{J_R} + z)^2 + (\bar{j} + z)^2 - 2(\widehat{m}_{J_R} + z)(\bar{j} + z),$$

which simplifies to

$$S_1 = \Big((\widehat{m}_{J_R} + z) - (\bar{j} + z)\Big)^2 = (\widehat{m}_{J_R} - \bar{j})^2. \tag{34}$$

Combining (32), (33), and (34) gives Equation (7).

REFERENCES

[1] Ittai Abraham, Omar Alonso, Vasilis Kandylas, Rajesh Patel, Steven Shelford, and Aleksandrs Slivkins. 2016. How many workers to ask? Adaptive exploration for collecting high quality labels. In *SIGIR 2016*. 473–482. DOI:http://dx.doi.org/10.1145/2911451.2911514 arxiv:1411.0149

[2] Omar Alonso, Daniel E. Rose, and Benjamin Stewart. 2008. Crowdsourcing for relevance evaluation. *SIGIR Forum* 42, 2 (2008), 9. DOI:http://dx.doi.org/10.1145/1480506.1480508

[3] Peter Bailey, Nick Craswell, Ian Soboroff, Arjen P. De Vries, and Emine Yilmaz. 2008. Relevance assessment: Are judges exchangeable and does it matter. In *SIGIR 2008*. 667–674. DOI:http://dx.doi.org/10.1145/1390334.1390447

[4] Ben Carterette and Ian Soboroff. 2010. The effect of assessor errors on IR system evaluation. In *SIGIR 2010*. 539–546. DOI: http://dx.doi.org/10.1145/1835449.1835540

[5] Gordon V. Cormack and Thomas R. Lynam. 2006. Statistical precision of information retrieval evaluation. In *SIGIR 2006*. 533. DOI: http://dx.doi.org/10.1145/1148170.1148262

[6] Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)* 28, 1 (1979), 20–28. DOI: http://dx.doi.org/10.2307/2346806

[7] Ofer Dekel and Ohad Shamir. 2009. Vox populi: Collecting high-quality labels from a crowd. *COLT 2009 Proceedings of the 22nd Annual Conference on Learning Theory* (2009). http://eprints.pascal-network.org/archive/00005406/

[8] D. Goldberg, A. Trotman, X. Wang, W. Min, and Wan Z. 2017. Drawing sound conclusions from noisy judgments. In *WWW 2017*. 307–314. DOI: http://dx.doi.org/10.1145/3038912.3052570

[9] Laura A. Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *SIGIR 2004*. 478–479. DOI: http://dx.doi.org/10.1145/1008992.1009079

[10] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of IR techniques. *TOIS* 20, 4 (2002), 422–446. DOI: http://dx.doi.org/10.1145/582415.582418

[11] Manas Joglekar, Hector Garcia-Molina, and Aditya Parameswaran. 2013. Evaluating the crowd with confidence. In *KDD 2013*. 686. DOI: http://dx.doi.org/10.1145/2487575.2487595 arxiv:arXiv:1411.6562v1

[12] Gabriella Kazai, Nick Craswell, Emine Yilmaz, and S. M. M Tahaghoghi. 2012. An analysis of systematic judging errors in information retrieval. *CIKM 2012* (2012), 105. DOI: http://dx.doi.org/10.1145/2396761.2396779

[13] Gabriella Kazai and Mounia Lalmas. 2006. Extended cumulated gain measures for the evaluation of content-oriented XML retrieval. *TOIS* 24, 4 (2006), 503–542. DOI: http://dx.doi.org/10.1145/1185877.1185883

[14] Gabriella Kazai and Imed Zitouni. 2016. Quality management in crowdsourcing using gold judges behavior. In *WSDM 2016*. 267–276. DOI: http://dx.doi.org/10.1145/2835776.2835835 arxiv:arXiv:1011.1669v3

[15] J. Richard Landis and Gary G. Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics* 33, 1 (1977), 159. DOI: http://dx.doi.org/10.2307/2529310

[16] Le Li and Mark D. Smucker. 2014. Tolerance of effectiveness measures to relevance judging errors. In *ECIR 2014*. 148–159. DOI: http://dx.doi.org/10.1007/978-3-319-06028-6_13

[17] N. A. Macmillan and C. D. Creelman. 2005. *Detection Theory: A User's Guide*. Lawrence Erlbaum Associates. https://books.google.com/books?id=EQLUGpgN0q8C

[18] Alistair Moffat and Justin Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *TOIS* 27, 1 (2008), 1–27. DOI: http://dx.doi.org/10.1145/1416950.1416952

[19] Rebecca J. Passonneau and Bob Carpenter. 2014. The benefits of a model of annotation. *Transactions of the Association for Comptuational Linguistics* 2 (2014), 311–326.

[20] Benjamin Piwowarski, Andrew Trotman, and Mounia Lalmas. 2008. Sound and complete relevance assessment for XML retrieval. *TOIS* 27 (2008), 1:1–1:37. DOI: http://dx.doi.org/10.1145/1416950.1416951

[21] Tetsuya Sakai. 2016. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *SIGIR 2016*. 5–14. DOI: http://dx.doi.org/10.1145/2911451.2911492

[22] Tetsuya Sakai. 2016. Topic set size design. *IRJ* 19, 3 (2016), 256–283. DOI: http://dx.doi.org/10.1007/s10791-015-9273-z

[23] M. Sanderson, F. Scholer, and A. Turpin. 2010. Relatively relevant: Assessor shift in document judgements. In *ADCS 2010*. 60–67. http://www.scopus.com/inward/record.url?eid=2-s2.0-84872873938

[24] F. Scholer, Andrew Turpin, and Mark Sanderson. 2011. Quantifying test collection quality based on the consistency of relevance judgements. In *SIGIR 2011*. 1063–1072. DOI: http://dx.doi.org/10.1145/2009916.2010057

[25] Mark D. Smucker, James Allan, and Ben Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *CIKM 2007*. 623–632. DOI: http://dx.doi.org/10.1145/1321440.1321528

[26] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. 2008. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Conference on Empirical Methods in Natural Language Processing*. 254–263. DOI: http://dx.doi.org/10.1.1.142.8286

[27] W. Tang and Matthew Lease. 2011. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval*. 36–41. https://www.ischool.utexas.edu/.

[28] Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *IP&M*. 36, 5 (2000), 697–716. DOI: http://dx.doi.org/10.1016/S0306-4573(00)00010-8

[29] Jeroen Vuurens, Arjen de Vries, and Carsten Eickhoff. 2011. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *ACM SIGIR Workshop on Crowdsourcing for Information Retrieval*. 21–26.

[30] Yining Wang, Liwei Wang, Yuanzhi Li, Di He, Wei Chen, and Tie-Yan Liu. 2013. A theoretical analysis of NDCG ranking measures. In *Proceedings of the 26th Annual Conference on Learning Theory*. 1–30. arxiv:1304.6480

[31] William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *TOIS* 28, 4 (2010), 1–38. DOI: http://dx.doi.org/10.1145/1852102.1852106

[32] Bo Zhao, Benjamin I. P. Rubinstein, Jim Gemmell, and Jiawei Han. 2012. A bayesian approach to discovering truth from conflicting sources for data integration. *VLDB* 5, 6 (2012), 550–561. DOI: http://dx.doi.org/10.14778/2168651.2168656 arxiv:1203.0058

[33] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments? In *SIGIR 1998*. 307–314. DOI: http://dx.doi.org/10.1145/290941.291014

Received May 2017; revised December 2017; accepted February 2018