The Impact of Judgment Variability on the Consistency of Offline Effectiveness Measures

LIDA RASHIDI, The University of Melbourne, Australia
JUSTIN ZOBEL, The University of Melbourne, Australia
ALISTAIR MOFFAT, The University of Melbourne, Australia

Measurement of the effectiveness of search engines is often based on use of relevance judgments. It is well known that judgments can be inconsistent between judges leading to discrepancies that potentially affect not only scores but also system relativities and confidence in the experimental outcomes. We take the perspective that the relevance judgments are an amalgam of perfect relevance assessments plus errors; making use of a model of systematic errors in binary relevance judgments that can be tuned to reflect the kind of judge that is being used, we explore the behavior of measures of effectiveness as error is introduced. Using a novel methodology in which we examine the distribution of "true" effectiveness measurements that could be underlying measurements based on sets of judgments that include error, we find that even moderate amounts of error can lead to conclusions such as orderings of systems that statistical tests report as significant but are nonetheless incorrect. Further, in these results the widely used recall-based measures AP and NDCG are notably more fragile in the presence of judgment error than is the utility-based measure RBP, but all the measures failed under even moderate error rates. We conclude that knowledge of likely error rates in judgments is critical to interpretation of experimental outcomes.

CCS Concepts: • **Information systems** → **Evaluation of retrieval results**; **Test collections**; **Relevance assessment**; **Retrieval effectiveness**.

Additional Key Words and Phrases: Evaluation, relevance assessment, significance testing

ACM Reference Format:
Lida Rashidi, Justin Zobel, and Alistair Moffat. 2023. The Impact of Judgment Variability on the Consistency of Offline Effectiveness Measures. *ACM Transactions on Information Systems* 1, 1, Article 1 (January 2023), 31 pages. https://doi.org/10.1145/nnnnnnn

1 INTRODUCTION

Research in information retrieval (IR) makes widespread use of offline evaluation of retrieval experiments [46]. In this methodology, there is a document collection and a set of queries, and, for all or some of the query–document pairs, there are also relevance judgments provided by humans. Retrieval systems are measured by their ability to permute the documents such that for each query the generated ordering approximates the relevance judgments associated with that topic; we refer to these orderings as *runs*. The effectiveness measurements observed over the set of queries can then be averaged to obtain a score for the retrieval system; if more than one system is being considered, these measurements can also provide the basis for some form of paired statistical test. That is, the

Authors' addresses: Lida Rashidi, rashidi.l@unimelb.edu.au, School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia; Justin Zobel, jzobel@unimelb.edu.au, School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia; Alistair Moffat, ammoffat@unimelb.edu.au, School of Computing and Information Systems, The University of Melbourne, Melbourne, Australia.

1046-8188/2023/1-ART1 $15.00
https://doi.org/10.1145/nnnnnnn

topic score measurements for a set of systems can be compared and assessed for properties such as an ordering of the systems or the statistical significance of the relative ordering of a pair of those systems.

In simplest form the judgments are binary ("yes", this document is relevant for the query, or "no", this document is not relevant) and the measurement is, in essence, a calculation of the extent to which the "yes" judgments are densely packed at the start of the run. We assume this form of experiment here. To quantify that notion of "top-weighted density" a variety of calculation formulae have been proposed; here we focus on two well-known options, average precision (AP) [8] and rank-biased precision (RBP) [37], but also consider normalized discounted cumulative gain (NDCG) [23], precision-at-k-documents-returned (P@k), and reciprocal rank (RR).

Our purpose in this paper is to examine the reliability of these measurements in the presence of errors in the judgments. The source of these errors are the human *judges* who provide the judgments as assessments of the relevance of a given document to a specific query. Each judge provides their opinion on the question of relevance, perhaps after formal training, and perhaps in the context of detailed written guidelines and a carefully curated set of examples. There might also be multiple judges examine each query–document pair, and a voting scheme employed to consolidate their independent opinions. But even so, it *is* a question of opinion, and so any given set of judgments should be regarded as being indicative rather then exact.

Our arguments are based on the assumption that "error" can be defined in an absolute sense, but we acknowledge that inconsistency in judgments can be viewed from a variety of perspectives. Indeed, even when the same judge views the same document for a second time – perhaps in a different context and after seeing different documents in the preceding moments – it cannot be guaranteed that they will react with the same opinion. However, the volumes of data involved, and the pressure to complete them to a tight schedule, mean that judges make actual errors, as well as making assessments that are inconsistent or debatable. Thus, in our terms, every set of relevance judgments is likely to contain some amount of error or *noise*, appearing at an unknown rate.

That is, it is useful to think of judgments as a collection of (experimental) observations made by judges, and therefore that they are an *observed* attribute; in this paper we call these the *judgoids*.[1] The judgoids for a test collection rest on the *true* relevance judgments for each query–document pair, which we call *attestments*. The attestments – also sometimes known as *ground truth*, or *gold standard* – are unknown, but for our purposes they can be assumed to exist.

When viewed from this perspective, each set of judgoids arises from a combination of the underlying set of attestments with the errors – each of which changes the judgment on a document – that arise in the course of judgment elicitation. It is well known that judges can be inconsistent or to some degree unreliable, with causes ranging from the challenge of making a binary decision about an ambiguous document to outright mistake. Additionally, different kinds of judge behave in different ways. Some judgments are collected from experts; others from representative individuals drawn from a population of users; others from crowd workers, whose characteristics, typically, are unknown, and whose sense of "pride in work" may be under-developed (or in proportion to the net hourly rate at which they are getting paid). Across these diverse type of judges, there is inherent noise and uncertainty, including not only random errors but systematic issues such as popular documents being perceived as relevant.

[1] We propose this coinage based on an imperfect analogy with the word *factoid*, a term that was introduced to describe a statement that is widely accepted as being true even though it is not, or that is accepted as being true even though its truth is as yet unknown.

A range of models have been proposed to describe these errors for binary relevance judgments, which we consider in Section 2. In our experimental design, which does not involve human re-examination of the documents, it is not possible to reverse errors once they have arisen, so we take the original judgments as correct – an assumption that we regard as reasonable because that is how these judgments are used, and also because doing so is not a confound in our methodology. The error models can then be used to artificially add (further) errors to a set of judgments, with different parameters capturing different aspects of the behavior of judges: for example, whether the judge is *generous* or *parsimonious* [12]; or, alternatively, *expert* or *uninformed* [4, 26, 63].

That is, it is possible to simulate assessor behavior by introducing random noise according to a biased model, where the bias is intended to reflect behavior. Given a set of attestments, such a model can be used as a *perturbation*, resulting in a corresponding set of judgoids. By repeatedly generating such judgoid sets it is possible to observe the variations in system measurement that can arise – in principle – for a given kind of judge. Indeed, this is, in some respects, an obvious experiment, given such a model. But one of our key arguments in this work is that this approach is potentially uninformative, in that being able to know the range of faulty measurements that can arise by perturbing a given true measurement is only helpful if in fact we also know what that true measurement was. In practice we do not know what the true measurement was; that is, we are not in possession of the underlying attestments.

In this paper we develop a complementary approach. We use standard test collections to run large numbers of distinct experiments, and for each experiment repeatedly use the judgment models to generate numerous sets of judgoids. We can then collect the judgoid-based measurements and observe the distribution of attestment-based results to which they might correspond. This approach requires that we assume that the provided judgments are true, but as noted above we argue that doing so is robust because our investigation is into the behavior of the effectiveness measures; we are not examining the true properties of the systems. The outcome is that, with our method, given an effectiveness measure, a measurement, and an estimate of error (or knowledge of the kind of judge that was used), it is possible to gauge what the true measurement may have been.

To make use of a simile, measurements based on perturbed judgments are like squinting through a dense haze: no single observation allows a clear view, but if a large enough number of slightly different viewpoints are explored and aggregated, the broad extent of the object beyond the wall of haze may be discerned. In our case, observing the likely distribution of measurements based on the underlying true judgments will not tell us what those true measurements were, but we may be able to see their range. Our experiments employing this framework make use of the effectiveness measures noted above and several sets of TREC data, and lead to a clear conclusion: even a modest amount of noise can notably unsettle results, with large changes in system ordering and reports of statistically significant system differences in the opposite order to those found with the attestments.

Of particular concern is that the recall-based measures AP and NDCG appear to be more sensitive to noise than recall-independent measures such as RBP and even sometimes P@10. There are settings at which the system orderings reported by AP and NDCG are indistinguishable from random, while RBP remains relatively unchanged. Moreover, AP and NDCG are more likely than RBP to incorrectly report statistical significance.

Alarmingly, however, all of the measures failed as error rates were increased to levels that were still below those reported in investigations of real judgments. These results have major implications, including that many conclusions based on test sets and measurements may well be incorrect.

Section 2 discusses batch evaluation processes including document pooling and judging, IR effectiveness metrics, and system measurement and comparison. That provides the context for models of judgment error, which are introduced in Section 3 and used in Section 4 to inform our new proposed methodology. Section 5 then provides the results of experimentation based on

multiple TREC collections, making use of the new methodology to explore the robustness of a set of common IR effectiveness measures, and building evidence in support of the broad summary we have provided here. Finally, Section 6 completes our presentation, summarizing our findings and noting directions for possible future work.

2 BACKGROUND

A range of factors affect the reliability and generality of batch evaluation in information retrieval. Here we consider some of these issues, as background to the work reported in Sections 3 and 4.

2.1 Creating and Employing Relevance Judgments

As noted in Section 1, conduct of an offline experiment requires three resources: a collection of documents, a set of queries with answers among those documents, and a set of relevance judgments, often referred to as *qrels*. A common protocol for developing the relevance judgments associated with a given topic is to take a set of representative retrieval systems, generate a *run* from each for that topic, and then *pool* the runs to some selected depth d, to form the union of the top-d sets. This can be done to a uniform pre-chosen depth d across each of the topics; or, if there is a judgment budget that must be complied with, to a depth that is variable on a per-topic basis, balancing the overall judgment budget equally across the topics. A range of other judgment strategies have also been proposed [16, 31, 32, 38] that in various ways seek to shape the pool of documents so as to achieve the greatest benefit, where benefit is often (but usually on an informal basis) equated with "finding the greatest number of relevant documents".

It is effectively impossible to create comprehensive judgments in which every document is assessed against every query – in all but inconsequential collections the cross-product of topics and queries is just too huge for this to be viable. Any given pooling strategy thus develops a tiny minority of the available judgments, and investigations have been carried out to determine the extent to which the collected judgments are comprehensive [47, 64]. In response, techniques for quantifying the effect of missing judgments uncertainty have been proposed [37], and recommendations provided in regard to experimental methodologies that measure and disclose the degree to which the available qrels are a fit for the experiment being reported on [41]. Other researchers have sought to develop techniques for scoring incomplete rankings that bypass, or infer values for, any unjudged documents [3, 7, 44]. In addition, comparisons of absolute and relative system orderings have been carried out, considering the extent to which incomplete judgments affect experimental outcomes [12, 59, 63].

2.2 Inter- and Intra-Assessor Disagreements

Regardless of the issues associated with judgment coverage, there are also other variables that potentially affect the quality of any judgments that are collected, including document characteristics, judgment conditions, information requirement statement, and assessor behavior [29].

Most of the studies mentioned so far have employed multiple assessor groups, for example, expert and non-expert, to calculate the inter-assessor consistency; in general they have argued that, even with non-negligible fractions of disagreement amongst assessors, the quality of the test collection and the relative ranking of runs is only moderately unaffected. For example, Voorhees [59] conducted a study involving several TREC collections that assessed the impact of NIST assessor judgment disagreements on relative system ordering when measured via AP, finding that the rank correlation of the runs remained high even as the groups of assessors were altered. Whether the ordering of the strongest systems remained consistent was not examined, nor was preservation or loss of significance, but the overall outcome was of similar results from different judgments.

This result had been anticipated by Lesk and Salton [29] in a much smaller experiment making use of a collection of 1268 documents, who attributed the stability of relative system orderings to three contributing factors: effectiveness measures report the average performance over topics; assessor disagreements tend to arise more with low-ranking documents than with high-ranking ones; and effectiveness is measured on the relative positions of the relevant and irrelevant documents. Voorhees [59] disagreed with those three suggested explanations for system stability, and reported that the larger size of TREC collections and the diversity of topics were the reasons. Voorhees also noted that, for queries with a small number of relevant documents, effectiveness metrics such as AP were unstable. Using a different collection, a similar experiment was conducted by Burgin [10], comparing four groups of assessors, again finding that the resultant system orderings were relatively consistent.

In another early study, Cleverdon [15] introduced random changes to the set of relevant documents, namely for a subset of topics a random number of irrelevant documents replaced the known relevant documents. Cleverdon compared the relative rankings of different indexing methods using four independent sets of relevance judgments and found that system rank order was largely unaffected. Lee and Kantor [28] also carried out an early study of judgment consistency; and Saracevic [49] surveys much of that early work.

More recently, Bailey et al. [4] compared three groups of judges, with a group of expert judges asked to judge a subset of the topics, while two other groups with decreased levels of expertise judged all of the topics. Bailey et al. found that on a per-topic basis system scores varied as the judging groups were switched; and that while per-topic score variability was common, average system score variations were smaller.

Scholer et al. [50] developed further experiments to explore intra-assessor consistency, asking for each judge how stable their judgments are over time. Scholer et al. looked at duplicates and assessor consistency, as well as analyzing the impact on system orderings arising from factors such as the time between assessments, the previous decision made by the assessor, and the ordering of the document sequence. They found that order of judgment was the most significant influence on the consistency of each person's judgments.

Overall, these various investigations suggest that experimental outcomes – at least to the extent of system orderings as measured via an unweighted correlation coefficient – tend to be insensitive to errors in the relevance judgments. Our detailed experiments in Section 5 using a top-weighted correlation coefficient suggests that such a view is only partially reliable.

2.3 Modeling Assessor Behavior

Webber et al. [63] experimented with a value they called *document meta-rank*, using it to build a model of assessor behavior. Document meta-ranks are computed using the popularity of the document, its rank in the returned results for the topic, and its relevance score. They then compared meta-ranks and assessor disagreement and found that, while there is only limited disagreement on the top-meta-ranking relevant documents for a topic, the assessor might overturn the judgment with a probability close to 50% for low meta-ranking documents.

Soboroff et al. [55] investigated the impact that random relevance assessments have on the system orderings. They generated pseudo-qrels, by using a normal distribution centered around the ratio of relevant to irrelevant documents, sampling randomly on a per-topic basis. Although overall system orderings remained stable, the best-performing runs were the most affected by this sampling. Soboroff et al. also explored shallow pooling (top 10 retrieved) to increase the chance of including a rare relevant document in the pseudo-qrels. The use of a shallow pool improved the rank correlation of the runs for several of the ad hoc tracks that were explored.

Carterette and Soboroff [12] enumerated several patterns of assessor behavior that could affect the quality of relevance judgments, namely assessor optimism, fatigue, patience, enthusiasm, and conditional dependence of an assessor's judgments on their previous decisions. Carterette and Soboroff implemented their model on the Million Query Track and generated new query relevance judgments per topic, where they reported that assessor models that underestimate the number of relevant documents generate more accurate system orderings.

Li and Smucker [30] suggest that assessors be modeled using two orthogonal variables, discrimination and bias. Bias models the assessor's liberal, neutral, or conservative judging behavior, while discrimination accounts for their ability to discriminate relevant documents from irrelevant. Li and Smucker simulated query relevance judgments using a range of values for their assessor behavior model, finding that AP is relatively resilient to change when measured by rank correlation, and that while deep measures such as AP and NDCG benefit from conservative relevance assessments, shallow measures such as P@10 require less conservative assessing behavior.

In more recent work, Ferrante et al. [17] use an unsupervised approach, where the ratio of relevant to irrelevant documents is not known to assessors, and a uniform distribution is used to select a document as being relevant or not. Ferrante et al. consider random, underestimating, and overestimating assessor behavior models. Their framework merges different performance measures based on the estimated accuracy of crowd workers, and results in more stable system orderings as well as a more accurate prediction of system scores. Ferrante et al. [18] go on to consider the question of relevance itself, exploring the consequences of modeling relevance via continuous-valued binomial variables, as a further broadening of the traditional binary relevance assumption.

We make use of both the Li and Smucker [30] and Webber et al. [63] models of assessor behavior, adopting them because of their probability-based formalisms in terms of how to achieve simulated disruptions that then translate directly into an implementation, and because of their plausible underlying models of assessor behavior.

2.4 Crowdsourcing of Judgments and Queries

The ability to make relevance assessments was at one time regarded as requiring reasonably good language and analytical skills [59]. More recently, the advent of crowd-sourcing platforms has allowed the role to be broadened, and for much wider pools of people to be considered, each of them doing a smaller volume of work, and with individual variance (or errors) smoothed by amalgamating the responses of multiple people for each query–document pair. That ease of access has, in turn, led to further studies in terms of judgment consistency and of assessor agreement, and in some studied the consequential effect it has on experimental outcomes; see, for example, Büttcher et al. [11], Vuurens and de Vries [61], Kazai et al. [24], Kazai et al. [25], Scholer et al. [51], Turpin et al. [56], Maddalena et al. [35], Kutlu et al. [27], and Han et al. [22].

Another area in which crowdsourcing has been useful is as a method for eliciting queries. Traditional offline methodologies make use of a single query associated with each topic; if bootstrapping is applied, samples with replacement are then drawn from that fixed set of "canonical' queries. However, different users are likely to generate different queries even for the same underlying information need [6, 9, 39, 40, 66]. Crowd workers can be used to create such *query variations* as a response to a supplied *backstory*, a process that has been used to add query variations to existing test collections [5] and also as an integral part of the collection formation process [33].

2.5 Statistical Testing and Bootstrapping

Another critical aspect of offline evaluation is the use of statistical *confidence testing*. With multiple topic scores typically being averaged to obtain a system score, it is expected (and reasonably so)

that a p value will be computed, as evidence of the likely strength of any claimed relativity. Sakai [45] has surveyed testing prevalence; and other work has considered the complex question of which test to make use of [13, 20, 54, 57, 58]. Following that advice, in our experiments here we make extensive use of the Student t-test, which has been found to be robust for small samples and reliable with regard to Type-I errors [57].

Another key statistical technique is that of *bootstrapping*. If a single set of n observations drawn from an unknown distribution is all that is available, an arbitrary number of further statistically indistinguishable n-samples can be formed by sampling with replacement from that known set, thereby allowing certain attributes of the overall distribution to be inferred from the accumulation of corresponding individual attributes [43]. For example, a confidence interval for the mean of the underlying distribution can be established by considering the properties of hundreds or thousands of such "with-replacement" n-samples. When n is large, the distribution of replication factors for each individual item can be approximated by $Poisson(k;1) = [(1/e)/k! \mid k \in \{0,1,2,3,\ldots\}] = [0.368, 0.368, 0.184, 0.061, 0.015, 0.003, \ldots]$, that is, with each original item appearing zero times in any given replicate with a probability of 0.368, appearing one time with an equal probability of 0.368, appearing twice with a probability of 0.184, and so on. Bootstrapping across a set of n topics in this way allows other sets of topics to be inferred, and statistical confidence to be established experimentally.

Other experimental work has made use of the fact that random assignment of documents across a set of partitions also generates replicate collections with identical statistical properties [60]; of the fact that collections can be split based on source, domain, and content type, to generate distinct sub-collections [48]; and of the observation that all of topic, system, and sub-collection can be combined into a model that allows multiple system comparisons [19]. Related work has created sub-collections (also known as "shards"), and studied the impact of topic-shard interaction on system performance [21].

Zobel and Rashidi [65] combined these ideas, introducing the notion of *collection bootstrapping*, complementing previous evaluations that had made use of randomized collection splitting. Each generated bootstrapped collection is "like" the seed collection, but not identical to it, with each document replicated a number of times given by $Poisson(k;1)$. Statistics gathered over a large number of bootstrapped collection replicates can then be used to infer statistics in regard to the original collection. In follow-up work, Rashidi et al. [42] introduced a controllable systematic bias to the random collection sampling process, and explored the extent to which slightly different collections resulted in different experimental outcomes.

3 SYNTHETIC JUDGMENT ERRORS

Our goal is to explore the behavior of effectiveness measures in the presence of judgment error. Doing so requires a source of judgments annotated according to their correctness.

A possible approach would be to gather real judgments and use independent, high-quality assessments to determine which were correct and which faulty. However, such an exercise would be extremely costly, as doing so would be even more laborious than collecting the corresponding initial judgments. Moreover, to test sensitivity of effectiveness measures, a great many such assessments would be required, so that the total effort would be expected to be a large multiple of that of a standard full TREC experiment. Such an approach would also have significant drawbacks. First, the volume of errors would be fixed, making it difficult to assess how the measures perform at different error rates. The approach of, say, deleting some wrong judgments with the goal of varying the error rate would introduce a synthetic element that undermined the goal of using real data. Second, it would lack generality, because the errors could be argued to be particular to the kind of collection, the kind of information need, and the competence of the assessors; and with multiple assessors

of varying competence there would be confounds to the results. Third, even with multiple sets of reasonable judgments it is still unlikely that there are grounds to regard one of the sets as perfect or definitive, as no assessor is fully objective.

Instead, our preferred approach is to take a set of carefully created judgments and to perturb it by introducing synthetic errors in a controlled way. This allows us the benefit of building on previous work that has established methodologies for generating errors that are parameterized to anticipate variations in assessor characteristics. In this section we first review two principled strategies for generating perturbed query relevance judgments (Sections 3.1 and 3.2), which are due to Li and Smucker [30] and Webber et al. [63] respectively. Both of these strategies are designed to simulate the types of errors that might occur in the process of gathering query relevance judgments from assessors; with Li and Smucker's method mimicking errors in crowd-generated assessments, and Webber et al.'s method mimicking the errors made by expert judges. We then explain (Section 3.3) how these strategies can be used to created perturbed sets of relevance judgments, thereby setting the scene for Section 4, which goes on to describe how these approaches can be employed to evaluate the robustness of offline effectiveness measures in the presence of judgment errors.

We make use of standard terminology: the *false negative rate*, denoted *FNR*, is the fraction of relevant documents that are judged to be non-relevant, and the *false positive rate*, *FPR*, is the fraction of non-relevant documents that are judged to be relevant. Similarly, the *true positive rate*, *TPR*, is the complement of the false negative rate, with $TPR = 1 - FNR$; and the *true negative rate* is defined as $TNR = 1 - FPR$. The simulation that is described below is parameterized by a combination of *TPR* and *FPR*.

The basis of the simulations is an existing collection of judgments, which we call *attestments*, that are assumed to be correct. For our purposes, the potential presence of errors amongst these judgments is irrelevant, as we are investigating how the properties of measurement vary as the attestments are mixed with errors. However, there is also merit in using real judgments as the attestments, because any method for generating faux attestments would rely on assumptions that could render the results less applicable to real retrieval experiments.

Given a set of attestments, the simulations proceed by flipping individual judgments from "true" to "false" or vice versa, yielding a set of *judgoids*. While the original pool of judgments for each query has a known number R of relevant documents, the value of R is not considered during the simulation; neither R or some assumed faux R' is used as a target. That is, each flip is considered independently of the total number of relevant documents.

Flipping of judgments will have varying effects on system scores, and relative system scores. In particular, the change in any score will depend on where in that run that any flipped documents appear; and the change in any score comparison will depend on the consistency of that positioning across the set of systems. For example, if a popular relevant document – one that is placed near the top of the ranking in many of the system runs – is flipped to irrelevant, it will probably affect most systems in similar ways and relativities between systems may be not much changed. On the other hand, an incorrect label for a less popular relevant document is more likely to affect high-scoring systems than it is to affect low-scoring ones, since, if only a few systems have retrieved that document, there is a bias in favor of those being the best systems. Conversely, flipping a less popular irrelevant document will likely affect the outcomes of systems with lower overall scores.

3.1 Simulation of Random Judgment Errors

The first strategy we describe was originally explored by Li and Smucker [30]; we refer to it as *random qrel flips*. The intuition behind this method is that crowd workers such as Amazon Mechanical Turkers are unlikely to be subject experts in the tasks that they are asked to assess,

and are also likely to be working at speed because they get paid by the item, not by the hour. Both factors can be argued as likely to lead to systematic random assessment errors.

In this strategy, assessor behavior is modeled in terms of true positive and false negative rates, *TPR* and *FPR*. Li and Smucker [30] use signal detection theory as a model for describing assessors' discrimination and bias, denoted as *disc* and *bias* respectively, and defined as:

$$disc = z(TPR) - z(FPR)$$
$$bias = \frac{-1}{2}\left(z(TPR) + z(FPR)\right),$$

where $z(\cdot)$ denotes the inverse of the normal distribution function, *disc* denotes the assessor's ability to discriminate between a relevant and irrelevant document, and *bias* denotes the extent to which the user's judgments are *liberal* or *conservative*. As *disc* increases, the reliability of the assessor is increased. In tension with that relationship, as *bias* increases the assessor becomes more reluctant to deem a document relevant, that is, more conservative in their willingness to make a positive judgment.

It is thus possible to express *TPR* and *FPR* in terms of the discrimination and the bias:

$$TPR = CDF\left(\frac{disc}{2} - bias\right)$$
$$FPR = CDF\left(\frac{-disc}{2} - bias\right),$$

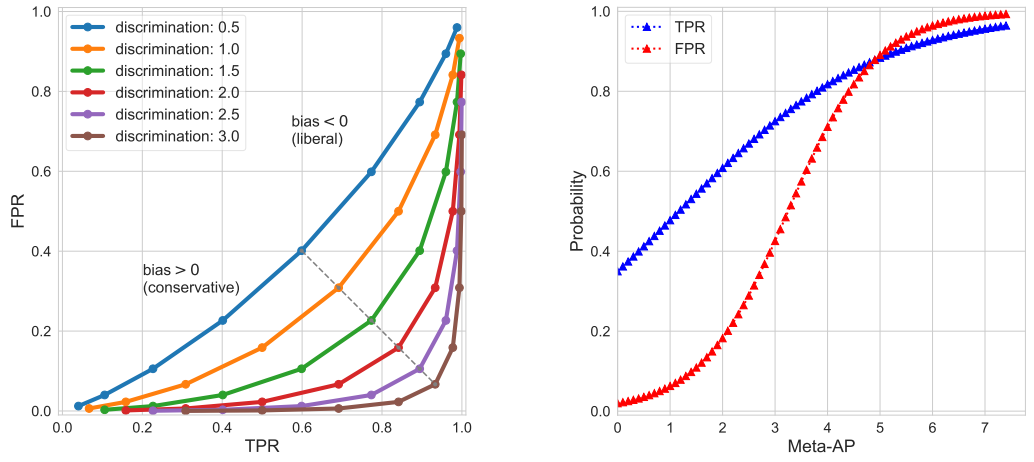
where $CDF(\cdot)$ is the cumulative density function of the standard normal distribution $\mathcal{N}(0, 1)$. That is, by varying *disc* and *bias* it is possible to simulate a range of assessor behaviors.

In the *random qrel flips* perturbation approach we assume that *TPR* and *FPR* are constant across all query–document pairs, and attributes of the assessment environment as a whole. The position in the run of each document is not considered, and the rate of flip is the same for all queries. Figure 1(a) shows the corresponding *TPR* and *FPR* values for assessors across a range of discrimination factors, $0.5 \le disc \le 3$; and a range of biases, $-2 \le bias \le 2$. Note that when $bias = 0$ the expected number of relevant documents remains the same after perturbation.

3.2 Simulation of Rank-Biased Judgment Errors

The second judgment perturbation strategy we consider is *rank-biased qrel flips*. This mechanism is designed to mimic the errors made by expert assessors, and proceeds from the assumption that rare relevant documents and popular irrelevant documents are the ones most likely to be misclassified. To model that propensity, Webber et al. [63] introduce the notion of the *meta-rank* of a document, which is an aggregation of its rank position in the set of runs generated by the available systems. As was the case with the work of Webber et al., the meta-rank method we use here is due to Aslam et al. [2], who suggest computation of *meta-AP* as a rank fusion technique calculated by averaging the contribution of each document in each run based on its rank position k in the run. The document's overall meta-AP score (which is only loosely connected to the metric AP) then reflects its normalized contribution across the set of runs.

The meta-AP of a document d is calculated as the average across the set of runs of that document's contribution scores, where the contribution score for a document at rank k in a run of length N is given by $1 + H_N - H_k$ if $k \le N$, and is zero otherwise; and where $H_k = \sum_{i=1}^{k}(1/i)$ is the k th harmonic number. For large N, we have $H_N \approx \gamma + \ln N$ where γ is a constant, meaning that $1+H_N-H_k \approx \ln(eN/k)$, and hence that the average of the document's contributions is proportional



(a) (b)

Fig. 1. (a) Expected error rates for the *random qrel flips* approach adapted from Li and Smucker [30], expressed as combinations of *TPR* and *FPR* and illustrated for a set of discrimination factors $0.5 \le disc \le 3$ and a range of liberal versus conservative bias factors *bias*. (b) The universal logistic regression model based on Webber et al.'s TREC 6 parameters for simulating assessor disagreements, expressed as *TPR* and *FPR* expectations as a function of document meta-rank as used in our *rank-biased qrel flips* simulation.

to the natural log of the geometric mean of that document's fractional positions across the set of runs, with fractional positions expressed as ratios relative to N. In the experiments described in Section 5 we typically employ $N = 1000$, which is the maximum run length in the ad hoc and robust tracks of TREC. As indicative values, the meta-AP of a document that is retrieved at rank 1 by all systems is 7.5 when $N = 1000$; if the same document is retrieved at rank 2 by all systems it gets a meta-AP of 7.0; and if a document is retrieved at rank 10 by half of the systems and not retrieved at all by the other half, it gets a meta-AP score of 2.8. Alternative run aggregation techniques could also be used to compute meta-ranks, and in other work have been used in rank fusion over query variations as well as over systems [6].

In analysis based on TREC datasets, Webber et al. [63] found that there is only limited disagreement between judges for relevant documents with a high meta-AP score, but that disagreements for relevant documents with low meta-AP scores approached 50%. Conversely, for irrelevant documents, low meta-rank documents tended to yield agreements between different assessors, while high meta-rank documents were often subject to disagreements. These observations suggest that, given a sequence of document ranks derived from a set of pooled runs, it is reasonable to use meta-rank as the basis of a model of the probability with which each document's relevance judgment might be flipped during the judgment elicitation process. More precisely, for expert assessors *TPR* and *TNR* are not distributed uniformly across the documents. Instead, *TPR* possesses a positive correlation with the aggregate rank of a document, and *TNR* has a negative correlation, the latter relationship then meaning that *FPR* also has a positive correlation.

Webber et al. [63] propose logistic regression models that estimate the probabilities of assessor agreement and disagreement for relevant and non-relevant documents, in each case as a function of meta-rank. They suppose that an assessor A has deemed a document d relevant or irrelevant, and estimate the probability p of a second assessor B also judging that document as relevant. Using

Table 1. Parameters used in conjunction with Equation 1 for assessor agreement estimation, taken from Webber et al. [63].

Attestment	To calculate	β_0	β_1
When $A = 1$	*TPR*, probability of a relevant staying relevant	−0.62	0.53
When $A = 0$	*FPR*, probability of a non-relevant being judged relevant	−3.90	1.20

the two corresponding pairs of parameters (β_0, β_1) shown in Table 1, the two required probabilities are modeled using two applications (one parameter pair (β_0, β_1) for $A = 1$, and then a second pair (β_0, β_1) for $A = 0$) of the fitted equation:

$$p(B = 1 \mid d, A) = \frac{e^{\beta_0 + \beta_1 \cdot \text{meta-AP}(d)}}{1 + e^{\beta_0 + \beta_1 \cdot \text{meta-AP}(d)}} . \tag{1}$$

That is, we take the attestments, that is, the original qrels, as being assessor A's decision, and seek to simulate assessor B via this model for each value for A. These two probabilities correspond to the *TPR* and *FPR* values in the *rank-biased qrel flips* strategy. The experiments described in Section 5 use Disks 4 and 5 from TREC; we accordingly make use of the parameters derived by Webber et al. [63] for the TREC 6 collection, as shown in Table 1. Figure 1(b) then shows how *TPR* and *FPR* vary, given the two logistic regression models. At the left-hand (low) end of the meta-AP range shown on the horizontal axis, relevant documents (the blue line) have a probability of less than 50% of retaining their positive label when judged by the second B assessor; while at the right-hand (high) end of the meta-AP range, non-relevant documents (the red line) are predicted to be almost certainly (incorrectly) judged as being relevant when considered by a second B assessor. Note that in the *rank-biased qrel flips* approach *TPR* and *FPR* are independently parameterized, so that – in contrast to *random qrel flips* and Figure 1(a) – it is not meaningful to plot them against each other.

Likely error rates on real data for both random and ranked models of error are discussed in Section 5.2.

3.3 Generation of Sets of Judgoids

Our interest is in how perturbed relevance judgments can affect system scores and system comparisons, and in how sensitive measurements are to the level of perturbation. We therefore require multiple sets of judgoids, which we generate by using the *random qrel flips* and *rank-biased qrel flips* approaches, with different discrimination and bias factors, which in turn lead to varying rates for *TPR* and *FPR*. To achieve any desired levels of *TPR* and *FPR*, we make use of the relationships noted by Li and Smucker [30] (see Section 3.1) connecting *disc* and *bias* on the one hand, and *TPR* and *FPR* on the other.

The overall process for generating perturbed query relevance judgments for each topic is shown by the pseudo-code in Figures 2 and 3. When the *random qrel flips* strategy is to be applied in Figure 3 the two sets $wgts_0$ and $wgts_1$ contain all-equal values; whereas if *rank-biased qrel flips* is to be applied, they contain the logistic-mapped (that is, via Equation 1) meta-AP scores of each document that has been judged for the selected topic, as a set of values between zero and one that indicate the relative preference for that document being in any of the generated subsets.

Each set of judgoids for a topic is then computed as follows. For a given level of discrimination and bias, steps 3–6 in Figure 2 first count the non-relevant judgments (set q_0) and the relevant judgments (set q_1); and then use the sizes of those two sets, plus the *disc* and *bias* parameters, to compute a false positive judgment count $n_{0\to1}$, and a true positive judgment count $n_{1\to1}$. Those two values are the respective target sizes for random subsets of q_0 and q_1, with that selection

1: **function** *create_judgoids*(*qrels*, *disc*, *bias*, *type*)
2: **input**: *qrels* is the initial set of relevance judgments for some topic, with $qrels[d] \in \{0, 1\}$ for each document d; *disc* and *bias* are the discrimination and bias parameters; and *type* is one of "*random qrel flips*" or "*rank-biased qrel flips*".
3: set $q_0 \leftarrow \{d \mid qrels[d] = 0\}$
4: set $q_1 \leftarrow \{d \mid qrels[d] = 1\}$
5: set $n_{0\to1} \leftarrow |q_0| \cdot CDF(-disc/2 - bias)$ ▷ number of negative judgments to be flipped, *FPR*
6: set $n_{1\to1} \leftarrow |q_1| \cdot CDF(disc/2 - bias)$ ▷ number of positive judgments to be retained, *TPR*
7: **if** *type* = "*random qrel flips*" **then** ▷ set the relative document weights
8: set $wgts_0[d] \leftarrow 1/|q_0|$ for all $d \in q_0$
9: set $wgts_1[d] \leftarrow 1/|q_1|$ for all $d \in q_1$
10: **else** // *type* = "*rank-biased qrel flips*"
11: set $wgts_0[d]$ from meta-AP(d) using Equation 1 and the parameters in row two of Table 1
12: set $wgts_1[d]$ from meta-AP(d) using Equation 1 and the parameters in row one of Table 1
13: set $q'_0 \leftarrow weighted\_subset(q_0, wgts_0, n_{0\to1})$ ▷ subset of "non-relevant" judgments to be flipped
14: set $q'_1 \leftarrow weighted\_subset(q_1, wgts_1, n_{1\to1})$ ▷ subset of "relevant" judgments to be retained
15: set $qrels' \leftarrow qrels$ ▷ start with the original *qrels*
16: **for** $d \in q'_0$ **do**
17: set $qrels'[d] \leftarrow 1$ ▷ judgment flip that contributes to false positive rate
18: **for** $d \in q_1 \setminus q'_1$ **do**
19: set $qrels'[d] \leftarrow 0$ ▷ judgment flip that detracts from true positive rate
20: **return** $qrels'$

Fig. 2. Generation of perturbed query relevance judgments for a single topic, given an initial reference set (*qrels*), and parameters for the judges' discrimination (*disc*) and bias (*bias*), which jointly determine the true positive rate (*TPR*) and the false positive rate (*FPR*). This algorithm makes use of the *weighted_subset*() function described in Figure 3, which constructs the subsets of judgments to be flipped.

process undertaken at steps 13–19. In between, two sets of weights are constructed, one containing a value between zero and one for each current non-relevant item, denoting its relative propensity to be flipped and thus become relevant; and one containing a value between zero and one for each current relevant item, similarly denoting its relative propensity to be retained as relevant. Those two *wgts* arrays are what drives the process shown in Figure 3, which uses them as guidance to bias the probabilities used to select a random non-uniform subset of its argument S.

Note that the two $wgts[]$ arrays are not probability distributions, and do not each sum to one – the values that they contain simply express relative emphases on the elements. Note also that under certain circumstances those emphases need to be moderated so that the target subset sizes ($n_{0\to1}$ elements from $wgts_0[]$, and $n_{1\to1}$ from $wgts_1[]$) can be achieved in expectation. The process shown in Figure 3 carries out that balancing, and then performs the subset generation. A key component is that the "sense" of the subset extraction is reversed if the relative ratios between the weights would suggest that any element must be over-sampled. If such a condition arises (the test at step 4) then the statements through to step 7 create a complement set based on reverse weights (the difference of each from 1.0), and then upon return the sense is flipped back, and the selected elements removed to make the desired subset.

Figure 4 provides two examples of how the selection is done, in both cases taking as input the same vector $wgts[]$ of $n = 9$ relative selection weights. In Figure 4(a) a subset of $n' = 2$ of the items is to be randomly selected. The computed *multiplier* (step 8 in Figure 3) is less than 1.0, and so

1: **function** *weighted_subset*(S, $wgts$, n')
2: **input**: S is a set of n items; $wgts$ contains n values that indicate the relative frequencies with which the corresponding items are to be selected into the subset, with $0 < wgts[i] < 1$; and $1 < n' < n$ is the target size of the extracted subset.
3: set $avg\_wgt \leftarrow (1/n) \cdot \sum\{wgts[d] \mid d \in S\}$
4: **if** $avg\_wgt < n'/n$ **then** ▷ a high fraction of S is being retained
5: set $wgts'[d] \leftarrow 1 - wgts[d]$ for $d \in S$ ▷ so compute the complementary set of weights,
6: set $S' = weighted\_subset(S, wgts', n - n')$ ▷ then determine a "not-retained" subset,
7: **return** $S \setminus S'$ ▷ and finally return its complement
8: set $multiplier \leftarrow n'/(n \cdot avg\_wgt)$ ▷ compute scale factor for weights, with $multiplier \le 1$
9: set $S' \leftarrow \{\}$
10: **for** $d \in S$ **do**
11: set $q \leftarrow wgts[d] \cdot multiplier$ ▷ q is desired expectation of item d in S'
12: **if** $random\_float(0.0, 1.0) < q$ **then** ▷ select d with probability q
13: set $S' \leftarrow S' \cup \{d\}$ ▷ and add it into S'
▷ S' now has expected size n'
14: **return** S'

Fig. 3. Creation of randomized weighted subsets of items, where the selection of items into the subsets is biased by the items' weights. For example, if $wgts[d_1] = 0.2$ and $wgts[d_2] = 0.3$, then d_2 is 1.5 times more likely to be placed into the subset S' as is d_1, but also subject to the overall ratio n'/n. For the *random qrel flips* strategy, all $wgts[d]$ values are equal. For the *rank-biased qrel flips* strategy, the $wgts[d]$ values are obtained through the application of Equation 1 to the set of meta-AP ranks for the elements $d \in S$, using selected parameters β_0 and β_1 (see Table 1).

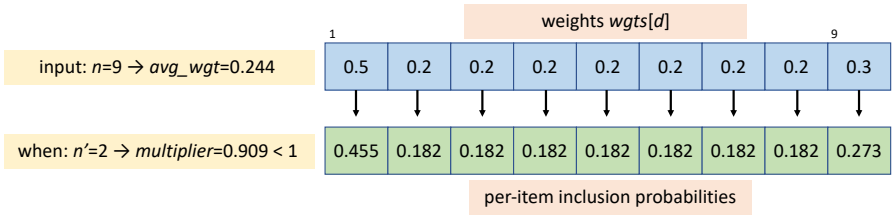
the operation can proceed using the per-item selection probabilities shown in the corresponding green cells, which sum to 2.0. With each of the $n = 9$ selection decisions an independent random event, summed across the $n = 9$ items there will thus be, in expectation, $n' = 2$ that are chosen. In Figure 4(b) a subset of size $n' = 7$ is sought. Now the computed *multiplier* is greater than 1, and per-item selection probabilities cannot be employed, because some might end up being greater than 1.0. To handle this situation, the vector of weights is complemented relative to the upper limit of 1.0 (step 5), and then the same process applied via a single recursive call (step 6) to determine a vector of *exclusion* probabilities (again, shown in green in the figure) of $n'' = n - n' = 2$ items to be removed (step 7), thereby forming the desired subset.

The random number generation process at step 12 of Figure 3 means that each time the function *weighted_subset*() is called a different non-uniform sample of the attestments is selected and either flipped when taken from q_0, or retained when taken from q_1. Any desired number of alternative sets of judgoids can thus be constructed by iterating the mechanism described by the pseudo-code shown in the two figures.

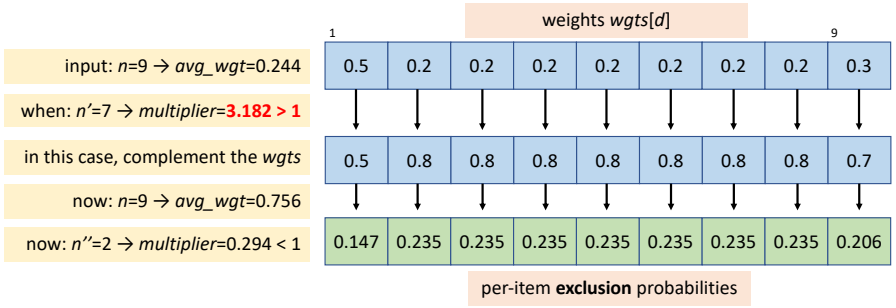
4 METHODOLOGY

For offline system evaluation in practice, the experimenter does not have the attestments but only a single set of judgoids, which are a mix of accurate judgments and errors. Had the experimenter been in possession of the attestments, they could observe the true result; however, all they have is an approximation of unknown characteristics.

Our interest is in what attestments might correspond to a given set of judgoids, and in particular how the judgoid-based results might vary from the results that the attestments would have produced. As the actual attestments are unknown (and unknowable), we instead must, as a substitute, explore



(a) per-item inclusion probabilities to attain an expected $n' = 2$ items from $n = 9$



(b) per-item exclusion probabilities to attain an expected $n' = 7$ items from $n = 9$

Fig. 4. Example showing the computations performed by Figure 3 for a set of $n = 9$ weights: (a) with $n' = 2$; and then (b) with $n' = 7$. In part (b) the high *multiplier* that is computed (the test at step 4 in Figure 3) means that the input weights first need to be complemented (step 5) so that per-item *exclusion* probabilities are instead computed.

plausible attestments. Thus the experimental protocol we employ is designed to demonstrate the distribution of possible "correct" results that might correspond to an observed result. With a perturbed result in hand, and an estimate of (or guess at) the likely true positive rate (*TPR*) and false positive rate (*FPR*), an indication of the true result can be obtained. Such a process is analogous to removal of noise from an image, or cleaning an audio signal.

Achieving this involves having a mapping from observed results to possible underlying true outcomes, as follows.

- Start with a set of high-quality judgments, which we treat as attestments.
- For each given combination of parameters, use the procedures described in Section 3.3 to repeatedly generate sets of judgoids, that is, perturbed judgments; and doing so systematically across a range of parameter settings.
- Use effectiveness measures to score "performance" according to each set of judgoids, giving a large number of observations for each combination of parameter values, as well as for the original attestments.
- Invert the observations, so that we obtain pairs of ⟨perturbed, original⟩ measurements, thus giving an *inverted* protocol that yields distributions of true measurements corresponding to any specific range of observed measurements.

This experimental protocol is illustrated in Figure 5. As discussed earlier, although we do not have access to attestments – indeed, they don't exist – for experimental purposes we regard a high-quality set of judgments as a reasonable substitute. That assumption allows us to generate perturbed judgments from a plausible starting point that can reasonably be assumed to resemble

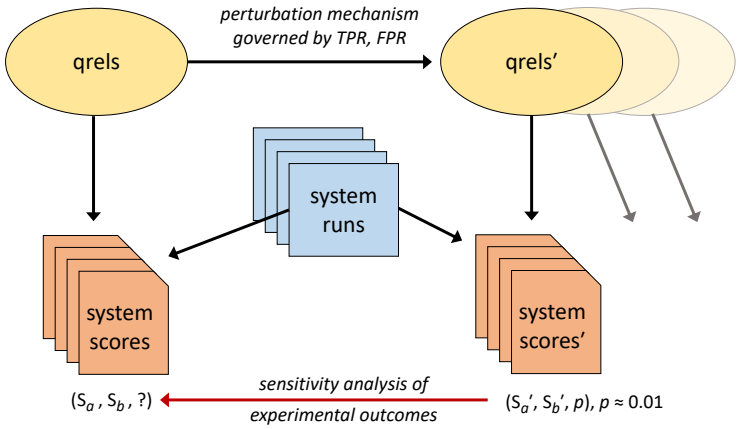


Fig. 5. Schematic layout of our experimental protocol, with multiple sets of perturbed qrels each used to evaluate and compare a set of systems, with each set of outcomes in a perturbed context compared with the corresponding outcomes in the original context.

true attestments, and then to observe the relationship between the original and the perturbed outcomes.

Explaining this further, we can expect sets of high-quality judgments to be similar in character to attestments, as they will be largely correct and have qualities that are reasonable approximations: the locations where the positive judgments occur in runs, the extent to which they distinguish between systems, and so on. These judgments can then be algorithmically degraded to produce large numbers of plausible sets of judgoids; we generate many sets of judgoids to smooth the effect of randomness, as any single set has some chance of being unrepresentative. With thousands of system-to-system comparisons on the (faux) attestments and each set of (faux) judgoids, we then have enough data to generate statistical observations of how the attestments and judgoids relate to each other.

Our interest is in enabling correct interpretation of practical experimental results, which are based on judgoids. For this reason we employ the inverted protocol shown in Figure 5. This reverses what might be seen as the typical or *forwards* structure of such experiments, in which a distribution of perturbed measurements is reported corresponding to some filtering operation in the original attestment-based measurement. The forwards structure is, we believe, unhelpful if the goal of the work is to enable richer understanding and interpretation of observed experimental results. What is instead needed is to be able to infer the distribution of actual "true" results that might correspond to an observation based on a set of judgoids, which is not possible if a forwards protocol is used. Hence our adoption of a *reverse protocol*: filtering by some characteristic according to the judgoids – which is where any incorrect conclusion, if arising, must have its evidential support – and then asking if the same conclusion would have been reached in the world of the attestments in the left side of the diagram.

Using reverse observations allows experimentation in a rich range of forms. For example, we could measure the performance of individual systems using the judgoids on the right side of Figure 5, and then ask how those outcomes relate to the "true" system scores in the context of the attestments. But more subtle comparisons are also possible, and the two specific situations that we explore in the next section are:

Table 2. Dataset statistics.

Collection	Systems	Documents	Topics	Avg. judged	Avg. rel.	Avg. irrel.
TREC7	103	358,493	50	1,607	94	1,513
Robust 2004 large	110	453,029	249	1,251	70	1,181

- We compare system orderings based on judgoids to those based on the attestments, seeking to identify how disruptive the various types of assessor error are to overall system versus system comparisons.
- We compare the outcomes of statistical tests of the significance of the difference in measured performance between two systems. Thus, for example, across a large number of sets of judgoids we can identify the system pairs whose performance is significantly different at the $p = 0.01$ level. Under the inverted protocol, we can then examine the distribution of the attestment-based p values for those same pairs of systems.

Specific experiments are explained in the next section. Note that when the runs produced by a large number of systems are available (together with the attestments that support their measurement) each true system average score might be mapped to hundreds of perturbed average system scores. Likewise, comparisons of 50 systems yields 1225 system pairs, and thus there maybe hundreds of thousands of perturbation-based p-values to examine.

5 EXPERIMENTS

Section 3 has described the *random qrel flips* and *rank-biased qrel flips* methods for generating judgoids from attestments; and Section 4 has presented our preferred *reverse* protocol for examining the relationships between perturbed and original results. This section builds on those ideas, presenting the experimental context and then exploring the properties that emerge under perturbation. In particular, the robustness of results as assessed via a range of standard effectiveness measures is of critical concern.

5.1 Collections and Measures

We make use of the binary judgments created for two TREC[2] data sets: TREC 7 (Disks 4 and 5) and TREC Robust 2004 (Disks 4 and 5, minus the congressional records).The dataset statistics are summarized in Table 2. These judgments, which are created to a high standard by the TREC processes, are used in our experiments as attestments. They undoubtedly do still contain errors, but of necessity are assumed to be a usable approximation of true attestments.

All 50 queries associated with TREC 7 were used. The Robust 2004 collection has a much large number of associated topics; these were divided into two sets, one with 50 queries and another with 199 queries, to allow consideration of topic set size upon experimental stability; in this collection, the mean judged relevant and irrelevant documents across the full set of 249 queries are 70 and 1,181 respectively. The sampling process used to obtain the two subsets was that every fifth query was assigned to the smaller query set. The runs from the top 50 systems in these two TREC rounds were selected, with "top" defined by system average for the metric rank-biased precision (RBP) [37] using a persistence value of $\phi = 0.95$, which corresponds to users who, on average, examine the first twenty documents in each run, and who value relevance in rank position 1 approximately 2.65 times more highly than they value relevance in rank position 20. As a check, we also selected the

[2]https://trec.nist.gov/

top 50 systems using AP and NDCG, and across the four combinations of metric and collection the minimum top-50 overlap against RBP was 46 out of 50.

We also used RBP (again with $\phi = 0.95$, a value that should be assumed throughout unless otherwise noted) as a metric with which to evaluate runs, along with several other well-known approaches: average precision (AP); precision at 10 (P@10); normalized discounted cumulative gain (NDCG) [23] with log base 2; and reciprocal rank (RR). Note that the use of $\phi = 0.95$ broadly corresponds to the expected evaluation depth attained by AP and NDCG [41].

Rank-biased overlap (RBO) [62] was employed in experiments that required comparison of system rankings. Rank-biased overlap is a top-weighted list similarity measurement that assigns larger penalties to differences at the head of a ranking than it does to differences that occur further down; we use a parameter of $\phi = 0.90$ except where otherwise specified, which corresponds to a probabilistic viewer of the rankings pair who on average scans from the top of the two runs, exits after looking at an average of 10 pairs of documents, and then assesses the degree of overlap they have observed. We mainly use RBO to compare system orderings, and hence evaluate it to depth 50, the number of systems employed in the experiments.

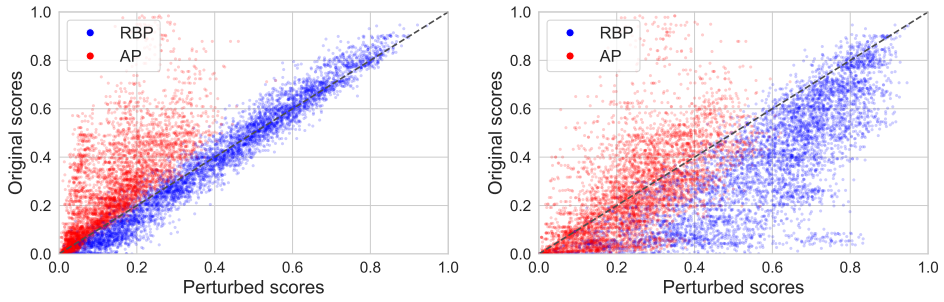
Rank-biased overlap is an overlap coefficient, and is zero only if the two lists are disjoint. When the two lists are permutations of each other there is an expected "background" RBO score that can be determined by a Monte Carlo technique that generates and scores a large number of random permutations. That process was used to establish the floor value applicable to these experiments: permutations of length 50, and a parameter of $\phi = 0.90$. The expected RBO for random orderings was found to be 0.194 ± 0.072 – that is, around 0.2, implying that reported RBO values of less than around 0.3 should be interpreted as meaning that the two orderings each containing 50 systems are uncorrelated. We also computed Kendall's tau in connection with system orderings, noting that it is not top-weighted.

5.2 Perturbed Judgment Sets

Each experiment – involving a single combination of parameters such as discrimination and bias – involved the generation of 100 perturbed qrels files, with each topic treated independently. The range of plausible values for discrimination and bias was borrowed from previous research; in particular, Li and Smucker [30] suggest varying the assessor's discrimination ability and bias in the ranges $disc \in [0.5, 3.0]$ and $bias \in [-3.0, 3.0]$ respectively. Smucker and Jethani [52] and Smucker and Jethani [53] had previously reported estimated average values for discrimination and bias of primary assessors (expert, or experienced) as $disc = 2.3$ and $bias = 0.37$ corresponding to $TPR = 0.78$ and $FPR = 0.06$; whereas secondary assessors (crowd workers) had less discrimination and were less conservative, with $disc = 1.9$ and $bias = 0.14$ corresponding to $TPR = 0.79$ and $FPR = 0.14$ respectively. Those estimates were made across 10 topics of the TREC 2005 robust track [52, 53].

In preliminary experiments we explored parameter settings close to these latter values, seeking to simulate crowd workers. However, we shifted to more conservative settings after we saw the results that arose – with $FPR = 0.15$ we observed results very close to random. That level of variability occurs because there are far more "not relevant" judgments in the attestments than there are "relevant" ones, and flipping even 10%–20% of those "not relevant" outcomes swamps the "relevant" attestments. The net effect is to shuffle the set of systems. Indeed, we felt that the quality of evaluations that arose with $FPR = 0.15$ would be regarded by readers as absurd. Experiments with $FPR = 0.10$ also in some instances showed large differences between the robustness of utility-based metrics and the frailty of recall-based metrics.

As a result we focus on the narrow range $FPR \leq 0.07$. Even so, there are clear measurement problems with these "expert level" judgments, as will be demonstrated shortly. We explored a range



(a) (b)

Fig. 6. Attestment-based system-topic metric scores plotted as a function of the corresponding judgoid-derived system-topic scores for the TREC 7 collection for a smart ($disc = 3.0$) neutral ($bias = 0$) assessor, using: (a) *random qrel flips*; and (b) *rank-biased qrel flips*. A random sample of 2% of the 250,000 measured effectiveness score pairs (50 systems, 50 topics, and 100 sets of judgoids) is plotted to illustrate the spread of the full distributions.

of *TPR* values but in many experiments report only $TPR = 0.93$, to allow direct comparison between results and between the behavior of ranked and random flipping of judgments.

5.3 Raw Score Changes

Figure 6 shows the run score perturbations that result from the two judgoid generation protocols. To form the two graphs the process shown in Figures 2 and 3 was applied to the TREC 7 qrels with a discriminating ($disc = 3.0$) and neutral ($bias = 0.0$) assessor assumed (that is, an "expert"), with 100 sets of judgoids generated. That resulted in a total of $50 \times 50 \times 100 = 250{,}000$ individual system–query runs (systems by topics by judgoids). A random sample of 5,000 (that is, 2%) of those runs was then taken, and judgoid-derived and attestment-derived metric scores for AP and RBP computed and scatter-plotted.

In the left pane, the *random qrel flips* process generates RBP scores (blue dots) that are visibly consistent between judgoids and attestments, and hence well correlated. In contrast, the AP scores (red dots) are more dispersed, and also show a clear pattern of shifting above the dotted mid-line, that is, of being numerically smaller with the perturbed judgments that with the attestments.

With the *rank-biased qrel flips* approach in the right pane, RBP scores tend to be below the mid-line; judgoid-based run scores tend to be larger than with the attestments, as popular non-relevant documents get flipped to relevant, a consequence of the biased subset selection employed in the *FPR* computation. In contrast, AP has many scores closer to the mid-line, but with a large mass of attestment scores that were near zero having much higher corresponding perturbed scores; that is, both of AP and RBP now show a high degree of dispersion.

However, the absolute run scores are in general not of direct interest, as their primary value is in system comparison; the key test is whether systems are affected in a consistent way by any pattern of drift in scores. The next two subsections examine what happens when the perturbed run scores are used in system-to-system comparisons.

5.4 Variations in System Rankings

In this experiment we explore how perturbed judgments affect system ordering, in which each system is ranked from 1 to 50 by its average measured score with ties (which are rare) broken at

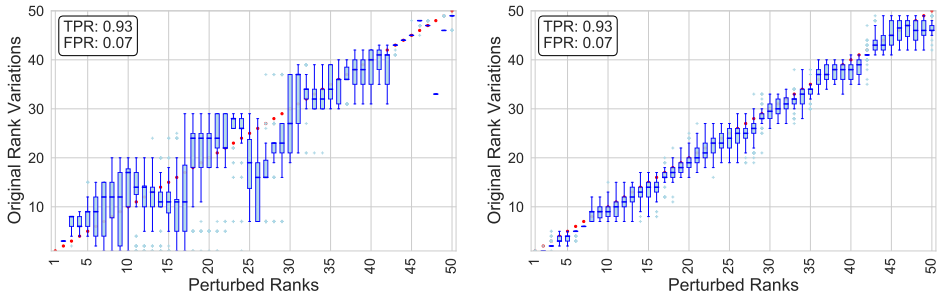


Fig. 7. Perturbed vs. original system ordering variations for a smart (*disc* = 3.0) neutral (*bias* = 0) assessor with *random qrel flips* on TREC 7 with 50 queries, 50 systems, and 100 judgoid sets, for: (a) AP (corresponding to mean RBO = 0.646); and (b) RBP (corresponding to mean RBO = 0.945).

random. Here, and in several of the other experiments discussed below, we report only results for AP and RBP, as well-performing and widely-used representative recall-based and utility-based metrics respectively.

For each of 100 sets of judgoids a system ranking is computed based on the mean metric score for each system across the corresponding set of topics, and then RBO is calculated, comparing that system ordering against the system ordering induced by the attestments. This process leads to 100 RBO scores for each parameter setting, each a measure of the similarity between results for one set of perturbed judgments and the original judgments. These scores are top-weighted, reflecting an interest in being able to identify a relatively small number of top-performing system. We then average the RBO values across the 100 sets of judgoids to obtain a single outcome value for each combination of parameters.

Before considering RBO scores, it is also helpful to look at rank ranges. These are computed out of the same experiment by tabulating, for each rank in each judgoid-induced system ordering, the corresponding system rank in the attestment-induced system ordering. This element of the *reverse protocol* yields information that helps answer the question "if a system is at rank x using the judgoids, where might it have been ranked using the underlying attestments". Figure 7 shows this type of output. A discriminating (*disc* = 3.0) and neutral (*bias* = 0.0) assessor is again assumed, the *random qrel flips* strategy is employed, and rank correspondences for AP are shown on the left and for RBP on the right. Each vertical box-whisker element shows the distribution of attestment rank positions corresponding to one judgoid rank position, with the solid area in each column depicting the middle two quartiles. The orange dots provide a guide that shows the location of the line of perfect consistency.

As is noted in the figure, these settings for *disc* and *bias* give a *TPR* of 0.93 and an *FPR* of 0.07. With around 1200 judged irrelevant documents per query, and 70 judged relevant, this corresponds to flipping to "off" around 4 previously positive judgments, and flipping to "on" around 85 previously negative judgments. Those error rates intuitively feel quite small; as noted earlier, we would generally be happy with such an accurate assessment from human judges. However, as the raw numbers show, they do mean that the incorrect positive judgments risk swamping the correct ones. Nonetheless, RBP appears to be robust, with a notable consistency apparent in the plot, and as a corroboration, a corresponding mean RBO of 0.945; whereas the corresponding RBO value for AP is 0.646. That lower correlation is clearly evident in the left plot in Figure 7, with a wide range of

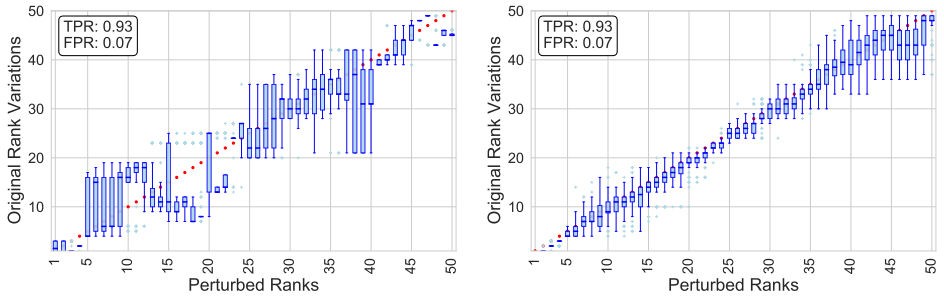


Fig. 8. Perturbed vs. original system ordering variations for a smart ($disc = 3.0$) neutral ($bias = 0$) assessor with *random qrel flips* on Robust 2004 with 199 queries, 50 systems, and 100 judgoid sets, for: (a) AP (corresponding to RBO = 0.742); and (b) RBP (corresponding to RBO = 0.940).

attestment-induced ranks corresponding to each judgoid-induced rank, even amongst systems that score well (the low ranks in the horizontal axis of the plot).

When the same experiment was carried out using the 199 Robust 2004 queries and the corresponding 50 best systems, similar results emerged, with RBO correlations for AP and RBP of 0.742 and 0.940 respectively. These results are shown in Figure 8, using the same presentation as in Figure 7. This is somewhat surprising, since using a larger set of queries appears to not be helpful in terms of convergence of results; and suggests that the number of queries is not an important factor. That is, for a given error rate, we tentatively hypothesize that the instability in ranks under AP with *random qrel flips* is innate, and isn't dampened by increasing the size of the query set. The reliance of AP on normalization by the number of relevant documents for each topic may be a contributing factor here, since each set of judgoids ends up with its own actual *FPR* rate, with the process in Figure 3 delivering perturbations in expectation, and not at a guaranteed rate that applies to every set of judgoids.

In other experiments with *random qrel flips* (results not shown here), RR and NDCG display a relatively wide range of variations in ordering similar to or worse than AP, while P@10 is more akin to RBP and is relatively stable.

Figure 9 presents the system rank relationships that arise when the *rank-biased qrel flips* approach is applied to the TREC 7 documents and queries, again for a smart neutral assessor. Now both AP and RBP show severe degradation in system ordering, presumably due to the fact that rank-biased perturbation particularly affects top-ranking systems – they get to the top by returning some unpopular yet relevant documents, which are exactly the ones that have a relatively high likelihood of being flipped and deemed "not relevant" in the judgoid sets. In particular, both measures have RBO scores of approximately 0.2; as discussed in Section 5.1, that means that the degree of judgment distortion deployed to obtain Figure 9 would render an experimental outcome essentially meaningless.

The *rank-biased qrel flips* perturbation method also affects outcomes for the Robust 2004 data and topics (not shown in a figure), but not as gravely, with RBO values of 0.517 and 0.629 for AP and RBP respectively. Analysis of the TREC 7 runs revealed a high level of diversity, and more distinctiveness between runs in terms of documents returned, including in terms of relevant documents returned. On the other hand, the Robust 2004 runs had more overlap with each other, and hence a lesser degree of vulnerability to the volatility introduced by *rank-biased qrel flips*. Even so, there was still

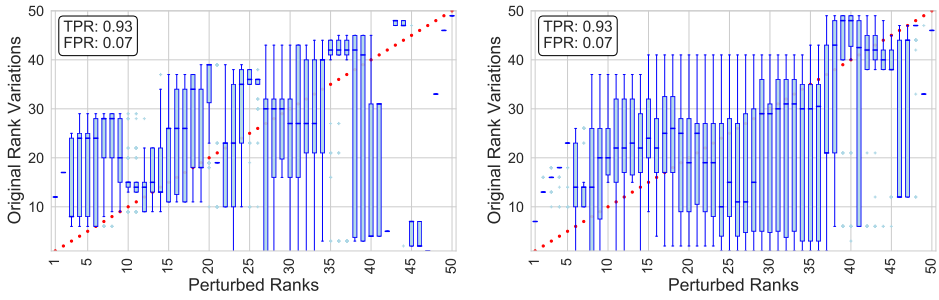


Fig. 9. Perturbed vs. original system ordering variations for a smart (*disc* = 3.0) neutral (*bias* = 0) assessor with *rank-biased qrel flips* on TREC 7 with 50 queries, 50 systems, and 100 judgoid sets, for: (a) AP (corresponding to RBO = 0.196); and (b) RBP (corresponding to RBO = 0.189).

wide variation in attestment rank for Robust 2004 systems that were top-scoring according to the judgoids, showing that even RBO scores of 0.5 or 0.6 are not necessarily a cause for celebration.

Table 3 presents a broader sweep of results, focusing on mean RBO scores between judgoid-induced system rankings and attestment-induced system rankings. Five effectiveness measures, two different true positive rates, and four different (and smaller) false positive rates are shown, for each of four different experimental configurations. As previously noted, RBO is calculated with a persistence of $\phi = 0.90$ to depth 50; with the greatest RBO score for each combination of collection, *TPR*, and *FPR* is highlighted in blue. Figure 10 shows the corresponding outcomes for the TREC 7 dataset. In Figure 10 the *TPR* = 0.9 results for each metric and for the four *FPR* values are laid down first via the bars in the solid colors, and then the *TPR* = 0.7 results are laid over via the hatching.

To confirm the validity of these results, we used a paired Student t-test to assess the significance of the differences between the results for RBP and AP. For *random qrel flips*, across the 24 cases in Table 3 and Figure 10 the computed p value was less than 10^{-6} in all but three instances, in which the difference was not significant. For *rank-biased qrel flips*, AP is significantly better than RBP in one case but RBP is significantly better than AP in the other 23, always with p similarly close to zero. We thus conclude that RBP does indeed lead to smaller RBO values when rankings induced by perturbed qrels are being compared to the rankings induced by the original attestments.

Given that the results shown in Table 3 and Figure 10 can be regarded as being significant, our overarching observations are as follows.

- In all cases, RBO falls with decreasing *TPR* and with increasing *FPR*. This demonstrates that increasing error rates do degrade the reliability of measurements – a completely unsurprising result, but nevertheless useful confirmation that the experimental framework is behaving as predicted.
- As already noted, the *rank-biased qrel flips* results for TREC 7 decrease quickly, a consequence of the distinctiveness of the systems used in that round of experimentation, and the large number of documents retrieved by the top systems that are not found by other systems.
- The *rank-biased qrel flips* protocol degrades rankings more than does the *random qrel flips* protocol, sometimes only slightly but sometimes substantially. As discussed above, this

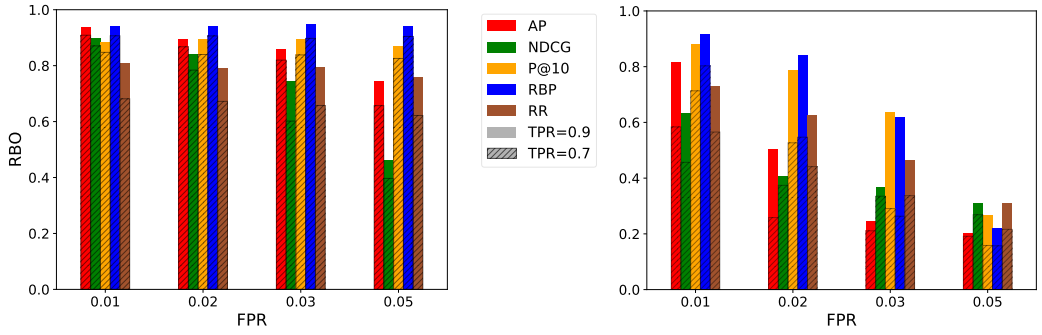
Table 3. Mean RBO across system orderings (100 values averaged for each data point) for original vs. perturbed qrels. As a reference, randomly permuted orderings of 50 items result in RBO scores of 0.194 ± 0.072. All RBO scores within 0.005 and 0.025 of the highest RBO value in each setting are shown in black and dark gray respectively.

Metric	*TPR* = 0.9, *FPR* shown below				*TPR* = 0.7, *FPR* shown below			
	0.01	0.02	0.03	0.05	0.01	0.02	0.03	0.05
Robust (50), *random qrel flips*								
AP	0.921	0.894	0.870	0.842	0.872	0.862	0.808	0.823
NDCG	0.893	0.866	0.848	0.812	0.862	0.847	0.810	0.767
P@10	0.855	0.838	0.840	0.816	0.744	0.710	0.694	0.702
RBP	0.928	0.926	0.929	0.921	0.888	0.881	0.874	0.873
RR	0.597	0.592	0.543	0.529	0.467	0.467	0.468	0.437
Robust (50), *rank-biased qrel flips*								
AP	0.896	0.854	0.834	0.766	0.851	0.822	0.800	0.699
NDCG	0.881	0.842	0.813	0.742	0.847	0.809	0.763	0.660
P@10	0.837	0.799	0.777	0.709	0.771	0.743	0.704	0.645
RBP	0.929	0.914	0.893	0.846	0.903	0.883	0.854	0.783
RR	0.546	0.488	0.423	0.378	0.486	0.436	0.396	0.344
Robust (199), *random qrel flips*								
AP	0.911	0.883	0.861	0.805	0.891	0.866	0.836	0.744
NDCG	0.948	0.917	0.878	0.737	0.929	0.888	0.818	0.697
P@10	0.910	0.913	0.899	0.882	0.816	0.818	0.807	0.785
RBP	0.964	0.963	0.954	0.945	0.926	0.931	0.917	0.894
RR	0.764	0.753	0.731	0.705	0.663	0.645	0.626	0.582
Robust (199), *rank-biased qrel flips*								
AP	0.870	0.709	0.612	0.532	0.754	0.633	0.574	0.503
NDCG	0.896	0.809	0.678	0.595	0.851	0.693	0.619	0.559
P@10	0.908	0.859	0.830	0.746	0.829	0.806	0.769	0.695
RBP	0.924	0.851	0.758	0.657	0.859	0.759	0.697	0.604
RR	0.733	0.675	0.621	0.555	0.663	0.622	0.543	0.490

occurs because *rank-biased qrel flips* has a disproportionate impact on the measurement of the top-scoring systems.[3]

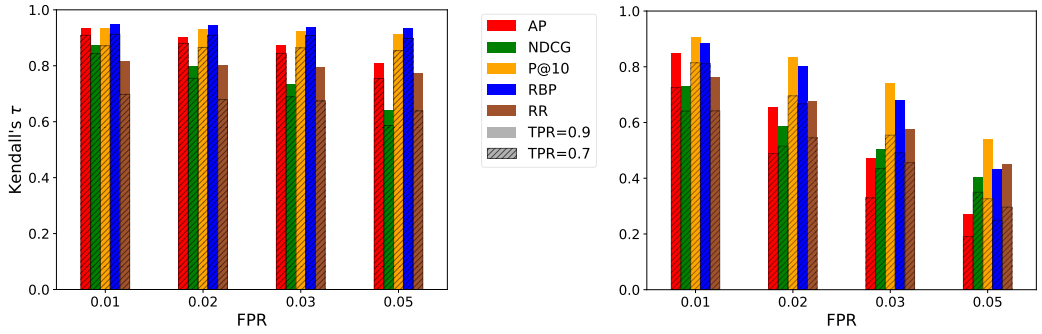
- The recall-based measures AP and NDCG are in all but one case more vulnerable to error than is the utility-based measure RBP. Reciprocal rank is significantly worse than the alternatives, consistent with other work has shown that it is insensitive and unstable [42].
- In terms of robustness of experiments in the presence of errors, increasing the volume of queries does not appear to help: the Robust 2004 (199) outcomes are no better than the outcomes with 50 queries.
- Rank-biased precision is less sensitive to error than are the other measures, with the exception of P@10 on the Robust 2004 (199) collection with *rank-biased qrel flips*, where RBP falls slightly behind. At low error rates it gives superior performance to the other measures, though the

[3]Thus demonstrating that consensus rank may be a good indicator of relevance but is a poor indicator of excellence.



(a) (b)

Fig. 10. Mean RBO across system ordering for original vs. perturbed rankings across different measurements with $FPR \in [0.01, 0.02, 0.03, 0.05]$ and $TPR = 0.9$ (background coloring) and $TPR = 0.7$ (hatched overlay), for: (a) TREC 7 with *random qrel flips*; and (b) TREC 7 with *rank-biased qrel flips*.



(a) (b)

Fig. 11. Mean Kendall's τ across system ordering for original vs. perturbed rankings across different measurements with $FPR \in [0.01, 0.02, 0.03, 0.05]$ and $TPR = 0.9$ (background coloring) and $TPR = 0.7$ (hatched overlay), for: (a) TREC 7 with *random qrel flips*; and (b) TREC 7 with *rank-biased qrel flips*.

alternatives are also competitive. Then, as error rates increase, it tends to remain better for longer.

As a confirmation, we also used Kendall's τ to compare system rankings, to provide a non-top weighted perspective. Figure 11 shows the same trend of ranking correlations as observed for the RBO scores, namely that utility-based metrics are more robust to judgment errors than are recall-based measures. We also observe again that the *rank-biased qrel flips* results in a faster onset of degradation.

To further explore the contrast between *random qrel flips* and *rank-biased qrel flips*, the collective behavior of the ten top-scoring systems was compared to the remaining lower-scoring systems, noting that when the RBO persistence factor is 0.9 (used in Table 3 and Figure 10) the top ten systems contribute the majority of the overlap calculation. We compared them to the remainder

by building a consensus run for each topic that combined the 50 systems, ordering documents by decreasing meta-AP score. We then computed an RBO score for each topic and for each system relative to that consensus run. Because we are now comparing rankings of documents (rather than rankings of system), a much deeper overlap was sought, and RBO with a persistence factor of $\phi = 0.98$ was employed.

For TREC 7 the average RBO ($\phi = 0.98$) score across all queries for the top 10 systems was 0.358, with the average over the other 40 systems being much higher, at 0.522. Investigating further, we found that the top systems in TREC 7 return documents that are relevant yet not popular; these documents do not gain a high meta-AP score and thus tend to be favored (Figure 1(b)) for flipping by the *rank-biased qrel flips* approach. In contrast, the Robust 2004 systems tend to behave similarly to each other, with slightly better RBO scores for the top-ranking systems compared to a consensus run, relative to the non-top systems. That is, the top systems and non-top systems retrieve similar proportions of popular and unpopular documents.

In results not included here, we also carried out experiments using the TREC-9 collection and relevance judgments. While those results showed up further patterns of behavior, in very broad terms they fell between the TREC-7 results and the Robust 2004 in terms of their sensitivity to disruption.

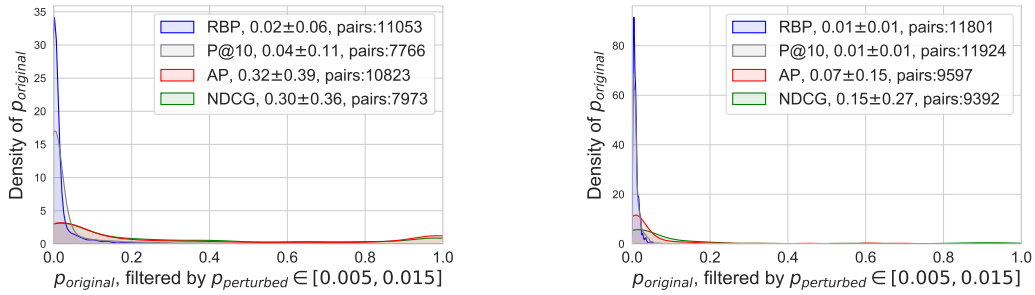
Our various results have several implications. First, they provide justification for the strategy of pooling; it is essential that all documents be considered equally, regardless of how many systems retrieved them. Second, it highlights the value of having a large number of systems contributing to the pool. Third, and most importantly, it suggests that popularity is only approximate as a predictor of assessor error. We have relied on it here, and are satisfied that our results are robust – as is illustrated by the strength of confirmation between random and ranked flipping of judgments – but it does imply that the model should be used with awareness of its limitations.

Arguably the most pertinent previous work to that reported in this section was that reported by Bailey et al. [4] and Voorhees [59]. In these papers, the authors investigated the agreement between different sets of relevance judgments. However, the different designs in this prior work mean that there are confounds to a comparison. Where we have been able to examine many millions of sets of synthetic judgments generated under a range of parameters, the prior work was of necessity limited to only one or two sets of additional real judgments on one or two test collections, on smaller numbers of systems and queries.

Bailey et al.'s results align well with ours, showing that substantial differences between systems can reverse when some of the judgments change. The methodology used by Voorhees makes direct comparison more difficult, as one of the sets of judgments is not based on pooling and thus many unjudged documents are assumed irrelevant in the score calculation, that is, there is an unknown error bound in the results, and the correlation function that was used (Kendall's tau) is not top-weighted. The greater volume of data in our work has allowed us to draw more conclusive inferences, with the limited scope of the previous experiments meaning that outcomes were, at least to some extent, tentative.

5.5 Stability of Significance Tests

A potential confound in the experiments reported in the previous subsection is that some disruptions to system orderings are uninteresting. When two systems have similar scores, or their differences in score are not statistically significant, changes in ordering may not be relevant to the hypotheses being tested. While we have noted consistent effects that vary monotonically in the degradation parameters, we nevertheless need to be alert to the possibility that what we have measured is the result of chance outcomes. To build confidence in the experiments reported in the previous subsection, we now restrict our attention to system differences that have been determined to be



(a) (b)

Fig. 12. Density plot of attestment p-values when pairs of systems are compared using the TREC 7 queries and attestments, with system pairs included in the tabulation if the *random qrel flips* p-value derived from a corresponding set of judgoids is in the range $[0.005, 0.015]$, for: (a) $TPR = 0.31$, $FPR = 0.07$; and (b) $TPR = 0.93$, $FPR = 0.07$. In this figure, an attestment p-value for a system pair $A > B$ is interpreted in the usual way when $p < 0.5$, and when $p > 0.5$ is plotted means that $A < B$ with 1-subtracted complement p-value used on the horizontal axis. That is, values that are close to 1 have the same significance meaning as values that are close to 0. Note that AP and NDCG result in almost identical curves in the left-hand graph; note also the different vertical scales in the two plots.

statistically significant. This is, after all, the core of a great many offline retrieval experiments: use of statistical significance tests to establish whether a new method can reasonably be claimed to be superior to alternatives. What we would hope to find is that when a pair of systems are significantly different when measured on a set of judgoids (which are all that is available in a typical experiment), then that significance would also be present in an evaluation based on the original attestments from which the judgoids were derived.

Consider the results shown in the two plots in Figure 12, which, as is the case throughout the rest of this subsection, concern the measures AP, NDCG, P@10, and RBP. The data set is TREC 7, with judgoids created via the *random qrel flips* protocol. From amongst the 125,500 individual comparisons (that is, covering all pairs from amongst 50 systems, with 100 sets of judgoids applied to each pair), we extracted all "System A, System B, judgoids" triples where the two systems were found to be significantly different (according to those judgoids) with a p value in the range $p \in [0.005, 0.015]$. Each of these outcomes is thus close to the "1 in 100" mistake level that would be regarded as being strong evidence of superiority. This filtering step is applied solely for the purposes of selecting for subsequent analysis a group of system pairs in which the relationship is highly significant and the p values are of comparable magnitude, and does not affect the "even more highly significant" outcomes derived when $p < 0.005$. The number of such outcomes for each of the four effectiveness metrics is shown in the legend; for example, on the left-hand side the number for RBP is 11,053.

We then took the matching "System A, System B" attestment-based p-values (as a multi-set) corresponding to the selected triples, and examined their distribution. The legend shows statistics of those four sets of corresponding p-values as computed via the attestments; on the left, for RBP it is 0.02 ± 0.06, for P@10 it is 0.04 ± 0.11, for AP it is 0.32 ± 0.39, and for NDCG it is 0.30 ± 0.36. The distribution of those attestment-based p-values is then shown by the inferred density curves.[4]

[4]We have used the Kernel Density Estimation utility from Python's seaborn package.

If system A outperforms system B on the judgoids, it is still possible for B to outperform A on the attestments, and indeed for B to be so much better than A that the results are statistically significant. To capture this possibility, in Figure 12 we report what we call *mapped* p-values:

(1) when $p > 0.5$, it indicates that the attestment and judgoid results disagree with $1 - p$ used as the value plotted on the horizontal axis.
(2) when $p < 0.5$, it indicates that the attestment and judgoid results agree with p used as the value plotted on the horizontal axis.

Thus, as a further subtlety in this experiment, p-values that are greater than 0.5 correspond to reversed system mean scores. If the perturbed observation was, say, that system A outperformed system B with $p = 0.01$, then an attestment observation of $p = 0.99$ would mean that B outperforms A with $p = 0.01$ – a strong contradiction of the finding that was inferred using the perturbed judgments.

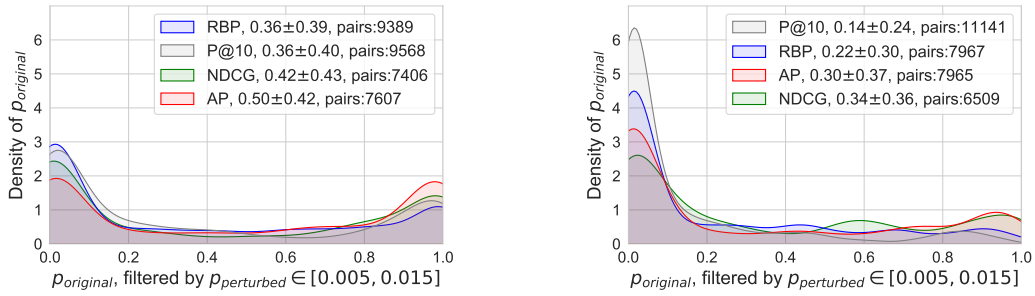
As can be seen, in Figure 12(a), which shows results with a very low *TPR* selected as an extreme case, the RBP p-values remain reasonably tightly grouped at the left, and show that it is behaving consistently: if the greatly perturbed judgoids suggest $p \approx 0.01$ for some A-versus-B system comparison, then the corresponding attestments also heavily favor small p-values for A-versus-B. On the other hand, the matched p-values for AP and NDCG are, more or less, spread right across the whole range. That is, in this context AP and NDCG have failed: the observed results in the presence of error bear little or no correspondence to the results that would have been obtained if the attestments had been available. This implies that experimental outcomes assessed with AP and NDCG – presuming that judgment errors of the supposed magnitude were indeed being made – would be consistently incorrect, even though each of the comparisons using the judgoids yielded statistical significance and high confidence.

Unfortunately, the same observations also apply to Figure 12(b), where the error rates are lower, and are at the "expert assessor" levels reported by previous authors. The distribution for AP and NDCG is not quite as poor, but it remains the case that, for most sets of judgoids and with the attestments taken as "truth", what are apparently successful findings using the judgoids are at best unsubstantiated, and at worst are false.

This finding has serious implications for use of AP and NDCG in practice. While RBP attestment results are centered on $p = 0.01$, in alignment with the judgoid results, those for AP and NDCG are centered on $p = 0.07$ and $p = 0.15$ respectively. That is, with even these low error rates the results show that a 1-in-100 chance of significance being a false positive corresponds to an underlying likelihood that the results are in fact not significant. Even a small volume of errors, well below that observed even among expert assessors let alone crowd workers, can lead to experimental outcomes that are simply wrong.

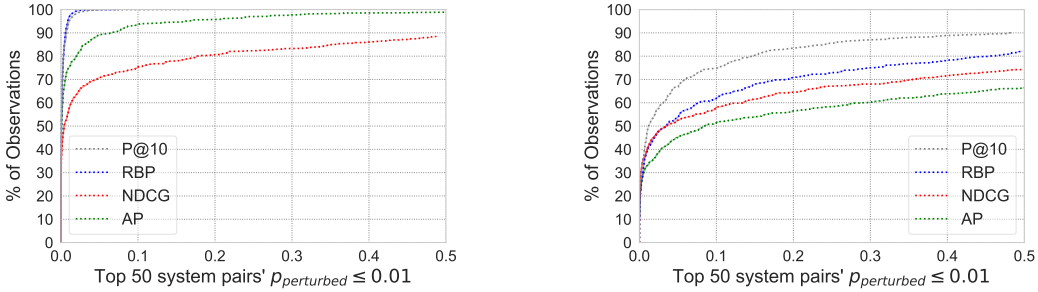
Experiments with *random qrel flips* on Robust 2004 (199) were very similar to those shown for TREC 7. A more interesting contrast is with the results for Robust 2004 (199) and *rank-biased qrel flips* perturbation, shown in Figure 13. (Note that in this figure the vertical scale is approximately one tenth that of the previous figure.) This data reveals failure for all four effectiveness measures. While RBP and P@10 are slightly better than AP and NDCG, in that they offer a higher peak on the left, the results show that the presence of rank-influenced judgment errors of the type suggested by Webber et al. [63] has the potential to render meaningless any assessment of statistical significance when comparing systems. The plot in Figure 13(b), with a reduced error rate, is only marginally better; judgoid significance centered on $p = 0.01$ corresponds to attestment significance of $p = 0.14$, $p = 0.22$ and $p = 0.34$ for P@10, RBP and NDCG respectively, even when the error rate is low.

A variant on the above results is shown in Figure 14. These two plots show the cumulative distribution of all attestment p-values that correspond (as a multi-set) to perturbed p-values using



(a) (b)

Fig. 13. Density plot of attestment p-values when pairs of systems are compared using the Robust 2004 (199) queries and attestments, with system pairs included in the tabulation if the *rank-biased qrel flips* p-value derived from a corresponding set of judgoids is in the range $[0.005, 0.015]$, for: (a) $TPR = 0.31$, $FPR = 0.07$; and (b) $TPR = 0.93$, $FPR = 0.07$. This figure uses the same conventions as Figure 12.



(a) (b)

Fig. 14. Cumulative distribution of attestment-derived p-values of pairs of systems, based on judgoid-derived experiments in which the p-value is ≤ 0.01 using TREC 7 with 50 queries and $TPR = 0.93$ and $FPR = 0.07$, for: (a) *random qrel flips*; and (b) *rank-biased qrel flips*.

judgoids that are ≤ 0.01. In other words, these are the "true" p-values for all of the "plausible experiment in the presence of judgment errors" observed p-values in which there was high statistical confidence in the measured outcome. The left plot shows results with the *random qrel flips* approach, the right shows results with the *rank-biased qrel flips* perturbation mechanism.

In these figures, the intercept with the right vertical axis (which is at $p = 0.5$) shows the fraction of system pairs that have the same ordering under both attestments and judgoids. For example, in Figure 14(b), for RBP, NDCG, and AP the proportions of system orderings that are consistent between judgoids and attestments are around 82%, 74% and 67% respectively. In Figure 14(a), effectively all of the p-values for attestments for RBP are well below 0.05 and over 90% are below 0.01, and the curve as a whole is in keeping with the expected false negative rate for a significance test. For the other effectiveness measures, the results are less consistent; NDCG is particularly poor, with less than 70% agreement at 0.05. Nor do the results improve much for lower error rates. It is clear that on this collection the recall-based measures are unacceptably vulnerable to errors in the judgments.

However, again echoing the earlier results, none of the metrics has done especially well on the Robust 2004 (199) collection, and nor do any of them do especially well in the face of *rank-biased qrel flips* perturbations. If indeed judgment errors are made according to the patterns modeled by Webber et al. [63], we may well need to find judges who are more expert than experts.

6 CONCLUSIONS

We have undertaken a systematic study of measurement reliability in information retrieval research based on two different methods by which relevance judgment errors might have arisen. Our goal was to explore the robustness of experimental conclusions that are based on sets of judgments that contain errors. As part of that investigation we have argued for an inverted protocol in which we ask the question, "if a result using perturbed judgments (the judgoids) is measured as being significant, what is the likelihood of the corresponding attestment-based outcome being in agreement?".

The results gathered for the *random qrel flips* strategy are moderately reassuring, in that low judgment error rates do not greatly affect system orderings and system-versus-system significance testing, at least when precision-based effectiveness such as metrics P@10 and RBP are employed.

But for other combinations our results provide clear warnings. With the *random qrel flips* perturbation policy and low error rates, for the recall-based metrics AP and NDCG, "significant" system-versus-system outcomes from judgoid-supported experiments are often not confirmed by the corresponding attestment-based comparisons. When the *rank-biased qrel flips* perturbation strategy is applied, not even the precision-based metrics can be regarded as reliable except when error rates are very low (note however that the effects of the error rates are not directly comparable between the two flipping strategies). Moreover, both perturbation strategies can lead to high levels of erroneous conclusions with all measures when error rates are increased; and we note that the levels of error considered in our experiments are lower than those estimated in literature that examined actual relevance judgments. We also found that RR was markedly less reliable than the other four metrics at almost all experimental settings.

Overall, the picture is deeply concerning. The *random qrel flips* perturbation model is hypothesized as resembling crowd-sourced judgments, with higher rates of both false positives and false negatives; whereas the *rank-biased qrel flips* perturbation model is hypothesized as applying to expert judges who have lower net error rates. But in both of those combinations our experiments have exposed trends that are worrying. Nor do we have a solution to this conundrum, except to reiterate that the recall-based metrics such as AP and NDCG tend to be more fragile in the presence of errors than are precision-based ones such as P@10 and RBP, at least in the contexts we have explored here.

As a final note, researchers and practitioners need to be aware that the fragility of the results is not detected by standard statistical techniques. An experiment using judgoids can yield a "near certain" outcome that is nevertheless the diametric opposite of the outcome inferred from the underlying attestments.

Limitations and Future Work

Our work here has made extensive use of TREC-based resources, primarily qrels and system runs in which (for the most part) each system executed the same single query in response to each topic. That means that the current judgment pools risk being uni-dimensional with respect to query variations [40]; broadening them to cover multiple queries for each underlying topic has been shown to be at least as expensive in terms of judgments as is broadening the set of systems involved [36]. Investigation of the sensitivity of metric scores, system orderings, and system-versus-system comparisons in the presence of query variations is an important next step.

Our methodology and experiments are designed for binary judgments; we do not have error models for multi-level judgments and thus cannot simulate them in a principled way. However, assessor disagreement and error is not limited to binary judgments and the evaluation measures have similar structures. It seems highly likely that the same kinds of issues will arise, but the sensitivity of the results to error, and indeed error rates, may be quite different in the multi-level case. Exploration of this issue would require perturbation models that are very different to those used here.

Note that our experiments are not designed to examine the relative validity of different methods of collecting judgments or the robustness of different kinds of judge. We used the same parameter settings for both methods of generating judgoids to ensure that the results were comparable, with the aim of showing that the same issues arises across different judgment methodologies.

Other researchers have considered the measurement reliability of even very shallow judgments when applied to a large number of topics [1, 14, 34]; this is another area where the reverse protocol that we have described might be an insightful way of exploring the dual areas of metric vulnerability and metric sensitivity in the face of judgment errors. However, it is already clear that current experiments have levels of vulnerability and sensitivity that risk undermining experimental outcomes, and which should be addressed if future research is to be robust.

ACKNOWLEDGMENTS

The third author was supported by grant DP190101113 from the Australian Research Council's Discovery Projects scheme.

REFERENCES

[1] N. Arabzadeh, A. Vtyurina, X. Yan, and C. L. A. Clarke. Shallow pooling for sparse labels. *Information Retrieval*, 25(4): 365–385, 2022.

[2] J. A. Aslam, V. Pavlu, and E. Yilmaz. Measure-based metasearch. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 571–572. ACM, 2005.

[3] J. A. Aslam, V. Pavlu, and E. Yilmaz. A statistical method for system evaluation using incomplete judgments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 541–548, 2006.

[4] P. Bailey, N. Craswell, I. Soboroff, P. Thomas, A. P. de Vries, and E. Yilmaz. Relevance assessment: Are judges exchangeable and does it matter. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 667–674, 2008.

[5] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. UQV: A test collection with query variability. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 725–728, 2016. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

[6] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. Retrieval consistency in the presence of query variations. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 395–404, 2017.

[7] C. Buckley and E. M. Voorhees. Retrieval evaluation with incomplete information. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 25–32, 2004.

[8] C. Buckley and E. M. Voorhees. Retrieval system evaluation. In E. M. Voorhees and D. K. Harman, editors, *TREC: Experiment and Evaluation in Information Retrieval*, chapter 3. The MIT Press, 2005.

[9] C. Buckley and J. Walz. The TREC-8 query track. In *Proc. Text Retrieval Conf. (TREC)*, 1999.

[10] R. Burgin. Variations in relevance judgments and the evaluation of retrieval performance. *Information Processing & Management*, 28(5):619–628, 1992.

[11] S. Büttcher, C. L. A. Clarke, P. C. K. Yeung, and I. Soboroff. Reliable information retrieval evaluation with incomplete and biased judgements. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 63–70, 2007.

[12] B. Carterette and I. Soboroff. The effect of assessor error on IR system evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 539–546. ACM, 2010.

[13] B. A. Carterette. Multiple testing in statistical analysis of systems-based information retrieval experiments. *ACM Trans. on Information Systems*, 30(1):4, 2012.

[14] C. L. A. Clarke, A. Vtyurina, and M. D. Smucker. Assessing top- preferences. *ACM Trans. on Information Systems*, 39(3): 33:1–33:21, 2021.

[15] C. W. Cleverdon. The effect of variations in relevance assessments in comparative experimental tests of index languages. Technical report, Cranfield University, 1970.
[16] G. V. Cormack, C. R. Palmer, and C. L. A. Clarke. Efficient construction of large test collections. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 282–289, 1998.
[17] M. Ferrante, N. Ferro, and M. Maistro. Aware: Exploiting evaluation measures to combine multiple assessors. *ACM Trans. on Information Systems*, 36(2):1–38, 2017.
[18] M. Ferrante, N. Ferro, and S. Pontarollo. Modelling randomness in relevance judgments and evaluation measures. In *Proc. European Conf. on Information Retrieval (ECIR)*, pages 197–209, 2018.
[19] N. Ferro and M. Sanderson. Sub-corpora impact on system effectiveness. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 901–904, 2017.
[20] N. Ferro and M. Sanderson. How do you test a test? A multifaceted examination of significance tests. In *Proc. Conf. on Web Search and Data Mining (WSDM)*, pages 280–288, 2022.
[21] N. Ferro, Y. Kim, and M. Sanderson. Using collection shards to study retrieval performance effect sizes. *ACM Trans. on Information Systems*, 37(3):30:1–30:40, 2019.
[22] L. Han, E. Maddalena, A. Checco, C. Sarasua, U. Gadiraju, K. Roitero, and G. Demartini. Crowd worker strategies in relevance judgment tasks. In *Proc. Conf. on Web Search and Data Mining (WSDM)*, pages 241–249, 2020.
[23] K. Järvelin and J. Kekäläinen. Cumulated gain-based evaluation of IR techniques. *ACM Trans. on Information Systems*, 20(4):422–446, 2002.
[24] G. Kazai, N. Craswell, E. Yilmaz, and S. M. M. Tahaghoghi. An analysis of systematic judging errors in information retrieval. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 105–114, 2012.
[25] G. Kazai, J. Kamps, and N. Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2013.
[26] K. A. Kinney, S. B. Huffman, and J. Zhai. How evaluator domain expertise affects search result relevance judgments. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 591–598, 2008.
[27] M. Kutlu, T. McDonnell, Y. Barkallah, T. Elsayed, and M. Lease. Crowd vs expert: What can relevance judgment rationales teach us about assessor disagreement? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 805–814, 2018.
[28] J. J. Lee and P. B. Kantor. A study of probabilistic information retrieval systems in the case of inconsistent expert judgments. *Journal of the American Society for Information Science*, 42(3):166–172, 1991.
[29] M. E. Lesk and G. Salton. Relevance assessments and retrieval system evaluation. *Information Storage and Retrieval*, 4 (4):343–359, 1968.
[30] L. Li and M. D. Smucker. Tolerance of effectiveness measures to relevance judging errors. In *Proc. European Conf. on Information Retrieval (ECIR)*, pages 148–159. Springer, 2014.
[31] A. Lipani, D. E. Losada, G. Zuccon, and M. Lupu. Fixed-cost pooling strategies. *IEEE Trans. on Knowledge and Data Engineering*, 33(4):1503–1522, 2021.
[32] D. E. Losada, J. Parapar, and A. Barreiro. Feeling lucky? Multi-armed bandits for ordering judgements in pooling-based evaluation. In *ACM Symp. on Applied Computing*, pages 1027–1034, 2016.
[33] J. Mackenzie, R. Benham, M. Petri, J. R. Trippas, J. S. Culpepper, and A. Moffat. CC-News-En: A large English news corpus. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 3077–3084, 2020.
[34] J. Mackenzie, M. Petri, and A. Moffat. A sensitivity analysis of the MSMARCO passage collection. arXiv:2112.03396, December 2021.
[35] E. Maddalena, M. Basaldella, D. De Nart, D. Degl'Innocenti, S. Mizzaro, and G. Demartini. Crowdsourcing relevance assessments: The unexpected benefits of limiting the time to judge. In *Proc. AAAI Conf. Human Computation and Crowdsourcing*, pages 129–138, 2016.
[36] A. Moffat. Judgment pool effects caused by query variations. In *Proc. Australasian Document Computing Symp. (ADCS)*, pages 65–68, 2016.
[37] A. Moffat and J. Zobel. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. on Information Systems*, 27(1):2.1–2.27, 2008.
[38] A. Moffat, W. Webber, and J. Zobel. Strategic system comparisons via targeted relevance judgments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 375–382, 2007.
[39] A. Moffat, F. Scholer, P. Thomas, and P. Bailey. Pooled evaluation over query variations: Users are as diverse as systems. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1759–1762, 2015.
[40] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. on Information Systems*, 35(3):24:1–24:38, 2017.
[41] A. Moffat, F. Scholer, and Z. Yang. Estimating measurement uncertainty for information retrieval effectiveness metrics. *ACM J. Data and Information Quality*, 10(4):10.1–10.22, October 2018.

[42] L. Rashidi, J. Zobel, and A. Moffat. Evaluating the predictivity of IR experiments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1667–1671, 2021.
[43] T. Sakai. Evaluating evaluation metrics based on the bootstrap. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 525–532, 2006.
[44] T. Sakai. Alternatives to bpref. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 71–78, 2007.
[45] T. Sakai. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 5–14. ACM, 2016.
[46] M. Sanderson. Test collection based evaluation of information retrieval systems. *Foundations & Trends in Information Retrieval*, 4(4):247–375, 2010.
[47] M. Sanderson and J. Zobel. Information retrieval system evaluation: Effort, sensitivity, and reliability. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 162–169. ACM, 2005.
[48] M. Sanderson, A. Turpin, Y. Zhang, and F. Scholer. Differences in effectiveness across sub-collections. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1965–1969. ACM, 2012.
[49] T. Saracevic. Effects of inconsistent relevance judgments on information retrieval test results: A historical perspective. *Libr. Trends*, 56(4):763–783, 2008.
[50] F. Scholer, A. Turpin, and M. Sanderson. Quantifying test collection quality based on the consistency of relevance judgements. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1063–1072. ACM, 2011.
[51] F. Scholer, E. Maddalena, S. Mizzaro, and A. Turpin. Magnitudes of relevance: Relevance judgements, magnitude estimation, and crowdsourcing. In *Proc. Wrkshp. Evaluating Information Access (EVIA)*, 2014.
[52] M. D. Smucker and C. P. Jethani. Measuring assessor accuracy: A comparison of NIST assessors and user study participants. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 1231–1232, 2011.
[53] M. D. Smucker and C. P. Jethani. The crowd vs the lab: A comparison of crowd-sourced and university laboratory participant behavior. In *Proc. SIGIR Workshp. Crowdsourcing for Information Retrieval*, pages 9–14, 2011.
[54] M. D. Smucker, J. Allan, and B. Carterette. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 623–632, 2007.
[55] I. Soboroff, C. Nicholas, and P. Cahan. Ranking retrieval systems without relevance judgments. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 66–73. ACM, 2001.
[56] A. Turpin, F. Scholer, S. Mizzaro, and E. Maddalena. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 565–574, 2015.
[57] J. Urbano. Test collection reliability: a study of bias and robustness to statistical assumptions via stochastic simulation. *Information Retrieval*, 19(3):313–350, 2016.
[58] J. Urbano, H. Lima, and A. Hanjalic. Statistical significance testing in information retrieval: An empirical analysis of type I, type II and type III errors. In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 505–514, 2019.
[59] E. M. Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management*, 36(5):697–716, 2000.
[60] E. M. Voorhees, D. Samarov, and I. Soboroff. Using replicates in information retrieval evaluation. *ACM Trans. on Information Systems*, 36(2):12:1–12:21, 2017.
[61] J. B. P. Vuurens and Arjen P. de Vries. Obtaining high-quality relevance judgments using crowdsourcing. *IEEE Internet Comput.*, 16(5):20–27, 2012.
[62] W. Webber, A. Moffat, and J. Zobel. A similarity measure for indefinite rankings. *ACM Transactions on Information Systems (TOIS)*, 28(4):1–38, 2010.
[63] W. Webber, P. Chandar, and B. Carterette. Alternative assessor disagreement and retrieval depth. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 125–134. ACM, 2012.
[64] J. Zobel. How reliable are the results of large-scale information retrieval experiments? In *Proc. ACM Int. Conf. on Research and Development in Information Retrieval (SIGIR)*, pages 307–314, 1998.
[65] J. Zobel and L. Rashidi. Corpus bootstrapping for assessment of the properties of effectiveness measures. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 1933–1952, Virtual Event, Ireland, 2020. ACM.
[66] G. Zuccon, J. Palotti, and A. Hanbury. Query variations and their effect on comparing information retrieval systems. In *Proc. ACM Int. Conf. on Information and Knowledge Management (CIKM)*, pages 691–700, 2016.

Received July 2022; revised February 2023