Estimating Measurement Uncertainty for Information Retrieval Effectiveness Metrics

ALISTAIR MOFFAT, The University of Melbourne, Australia
FALK SCHOLER, RMIT University, Australia
ZIYING YANG, The University of Melbourne, Australia

One typical way of building test collections for offline measurement of information retrieval systems is to pool the ranked outputs of different systems down to some chosen depth d, and then form relevance judgments for those documents only. Non-pooled documents – ones that did not appear in the top-d sets of any of the contributing systems – are then deemed to be non-relevant for the purposes of evaluating the relative behavior of the systems. In this paper we use RBP-derived residuals to re-examine the reliability of that process. By fitting the RBP parameter ϕ to maximize similarity between AP- and NDCG-induced system rankings on the one hand, and RBP- induced rankings on the other, an estimate can be made as to the potential score uncertainty associated with those two recall-based metrics. We then consider the effect that residual size – as an indicator of possible measurement uncertainty in utility-based metrics – has in connection with recall-based metrics, by computing the effect of increasing pool sizes, and examining the trends that arise in terms of both metric score and system separability using standard statistical tests. The experimental results show that the confidence levels expressed via the p-values generated by statistical tests are only weakly connected to the size of the residual and to the degree of measurement uncertainty caused by the presence of unjudged documents. Statistical confidence estimates are, however, largely consistent as pooling depths are altered. We therefore recommend that all such experimental results should report, in addition to the outcomes of statistical significance tests, the residual measurements generated by a suitably-matched weighted-precision metric, to give a clear indication of measurement uncertainty that arises due to the presence of unjudged documents in test collections with finite pooled judgments.

CCS Concepts: • **Information systems** → **Test collections**; **Retrieval effectiveness**; *Presentation of retrieval results*;

Additional Key Words and Phrases: Evaluation, test collection, effectiveness metric, statistical test, evaluation, information retrieval

1 INTRODUCTION

The effectiveness of document retrieval systems is commonly analyzed and compared via batch (or *offline*) evaluation. Batch evaluation approaches make use of three resources: a set of documents (the *collection*) felt to be a representative subset of the larger search context; a set of topics (detailed *information need statements*, or terse bag-of-word *queries*) felt to be a representative subset of the larger search context; and a set of *relevance judgments* (referred to as *qrels*, for short) which indicate

Authors addresses: Alistair Moffat and Ziying Yang: School of Computing and Information Systems, The University of Melbourne, Victoria 3010, Australia; Falk Scholer: School of Computer Science and Information Technology, RMIT University, Melbourne 3001, Australia.

1936-1955/2018/1-ART1 $15.00
https://doi.org/10.1145/3239572

which of the documents are relevant to which of the queries [33]. Relevance status is determined by human assessors, and is typically measured using an ordinal scale with two ("binary") or more ("graded") relevance levels.

When the collection is large, it is all but impossible to form comprehensive judgments, and normally only a subset of the documents are judged for each of the topics. One common way for a subset to be identified is via a process known as *pooling*, where a number of separate (and possibly also independent) retrieval systems all execute the queries, and the union of their top-d answer sets is formed, for some suitable value of d. For example, in the NIST-sponsored TREC-8 experiments carried out in 1999, a total of 129 systems were involved (with some research groups submitting multiple systems), pools to depth $d = 100$ were formed using a subset of 71 of those runs relative to a set of 50 topics, and a total of 86,830 judgments were carried out [41]. Of that average of 1736.6 judgments made per topic, an average of 94.6 documents per topic were deemed relevant by the NIST assessors, or around 5.4% of judged documents.

The relevance judgments formed by pooling are then used as an input to one or more *effectiveness metrics*, mechanisms that take a ranked list of documents and a qrels file, and compute a numeric score that indicates the relative quality of that ranking. The critical expectation is that rankings that are "good" will receive high scores; rankings that are "bad" will receive low scores; and hence that systems can compared based on their average scores, or based on the use of a statistical test in regard to their computed scores across the set of topics. But such comparisons between systems are vulnerable to a number of possible confounds, including whether or not the chosen effectiveness metrics correspond to attributes of the rankings that are observable by the users of the retrieval system and hence correspond to user satisfaction; whether or not the process for eliciting judgments is stable and consistent; and so on.

The usual IR batch evaluation framework gives rise to a number of issues that have potential implications for experimental reproducibility. These include the extent to which the same experimental outcomes would arise if more (or less) effort was allocated to the task of judging documents, by increasing (or decreasing) the pool depth; the extent to which the computed metric scores might vary as a result of the inevitability of them being calculated based on incomplete judgments; and the extent to which metrics can be regarded as being substitutable for each other when carrying out system evaluations. In particular, our investigation in this paper revisits the question of whether or not the pooling process yields relevance judgments that are sufficiently comprehensive to allow recall-based metrics such as average precision (AP) and normalized discounted cumulative gain (NDCG) to be accurately computed. Such metrics are usually regarded as being "deep", with influence accruing from a relatively high number of documents in each ranking. Estimating the reliability of such evaluations is an area of investigation that has a long history, discussed in more detail in Section 2. The lens we employ here to shed new light on this question is that of *residuals*, the fraction of the metric weight that is associated with unjudged documents when a weighted-precision metric such as RBP (rank-biased precision) [23] is used to score the rankings. It is not possible to compute residuals for recall-based metrics such as AP and NDCG directly, because they are not monotonically bounded in the presence of uncertainty. That is, as additional documents are judged the score of such metrics may increase or decrease [23]. However, it *is* possible to ask a two-part question: (1) which value (or values) of the RBP parameter ϕ yield system orderings closest to the system ordering associated with AP and/or NDCG, and how close are those system orderings; and then, (2) how big are the RBP residuals when that value of ϕ is used. That is, we estimate the residuals, or maximal score uncertainty ranges, associated with deep recall-based metrics, via a "best match" RBP parameter.

To create varying residual levels and hence varying levels of measurement uncertainty, we construct shallow pools as subsets of the standard relevance judgments, and compute the effect

increased uncertainty has on both metric score and on system separability via standard statistical tests. Our results show that the p-values associated with (for example) the Student t-test are largely uncorrelated with measurement uncertainty as represented by weighted-precision residuals, meaning that they are also relatively unaffected by the pooling depth used to form the judgments. That is, we demonstrate that system comparisons are relatively robust to the pooling depth involved in the experiment, even though the exact metric scores that are being compared might be subject to non-trivial uncertainty. In particular, we observe that low p-values can be regarded as reliable evidence of demonstrated relative system performance differences, but that if metric score values are to be considered as absolute rather than relative values, care should be taken to ensure that the corresponding residuals are commensurate with the degree of precision required when expressing those scores.

2 EFFECTIVENESS METRICS

We now summarize a number of topics that form the background of our experimental evaluation.

2.1 Effectiveness measurement in ranked lists

A very wide range of effectiveness metrics have been proposed for assigning a single numeric score to a ranked list of judged documents. Traditional set-based metrics such precision (the fraction of documents retrieved that are relevant) and recall (the fraction of relevant documents retrieved) have fallen out of favor, with top-weighted mechanisms that are better suited to ranked sets now preferred. Two broad classes have emerged, those that are *recall-based*, and those that are *utility-based*; see Moffat [19] for more in regard to these categories, and for a set of seven orthogonal properties that allow the contrasting behaviors of different metrics to be considered.

Dominant among the first recall-based group of metrics are *average precision* (AP) and *normalized discounted cumulative gain* (NDCG) [12]. Sakai [26] describes another recall-based metric, the Q-Measure, which is a weighted blend of AP and NDCG; and R-Prec, the precision at depth R, where R is the number of relevant documents for that topic, is a fourth recall-based metric. Moffat [19] brings together details of how all of these are computed.

In the second category, the utility-based group, there are similarly a range of metrics. Rank-biased precision (RBP) is one typical example of this genre [23]. Given a *persistence parameter*, denoted here as ϕ, RBP is computed as a weighted sum of relevance at ranks. In particular, if the relevance ranking is an ordered list $\mathbf{r} = \langle r_1, r_2, r_3, \ldots \rangle$, with $0 \le r_i \le 1$ the (binary or fractional-valued) relevance associated with the document at depth i in the ranking, then in an ideal sense,

$$\mathrm{RBP}(\mathbf{r}, \phi) = (1-\phi) \cdot \sum_{i=1}^{\infty} r_i \cdot \phi^{(i-1)} \,. \tag{1}$$

This definition assumes that relevance values are known for all documents, which is impractical. However, a key attribute of RBP – and all other weighted-precision metrics – is that when the judgments are incomplete, it is possible to compute a *residual*, the sum of the weights associated with unjudged documents. If J is the set of ranks at which documents have been judged, then the ideal computation of Equation 1 is replaced by

$$\mathrm{RBP}(\mathbf{r}, \phi, J) = (1-\phi) \cdot \sum_{i \in J} r_i \cdot \phi^{(i-1)} \,, \tag{2}$$

and the residual is computed as

$$\mathrm{RBPres}(\mathbf{r}, \phi, J) = (1-\phi) \cdot \sum_{i \notin J} \phi^{(i-1)} \,. \tag{3}$$

We will make extensive use of residuals as a way of quantifying the measurement uncertainty. For example, if some set of judgments J and relevance sequence $\mathbf{r}$ are such that $\text{RBP}(\mathbf{r}, \phi) = 0.2$ and $\text{RBPres}(\mathbf{r}, \phi) = 0.001$, the score of 0.2 is relatively "final" and cannot shift by much if more of the documents were to be judged. On the other hand, if $\text{RBPres}(\mathbf{r}, \phi) = 0.1$, then care needs to be taken when interpreting the corresponding RBP score – it might become substantially larger than 0.2 when more judgments are carried out.

The residual is the maximal additional score that could be achieved if every unjudged document $i \notin J$ was fully relevant and had $r_i = 1$, and is a bounding range rather than a confidence interval. If X is the true, or "full knowledge" RBP value according to Equation 1 and assuming that the relevance ranking is completely defined, then Equations 2 and 3 can be used to bound X when the judgments are incomplete:

$$\text{RBP}(\mathbf{r}, \phi, J) \le X \le \text{RBP}(\mathbf{r}, \phi, J) + \text{RBPres}(\mathbf{r}, \phi, J)\,.$$

In particular, if $J = \{1, 2, \ldots, d\}$ as a result of pooled-to-d relevance judgments, then the properties of the geometric sequence provide an upper bound on the RBP residual: $\text{RBPres}(\mathbf{r}, \phi, J) \le \phi^d$ [23].

Other utility-based metrics of interest are *reciprocal rank* (RR), *expected reciprocal rank* (ERR) [9], and precision itself. Moffat et al. [20] describe further weighted-precision metrics and the assumptions that they correspond to in terms of a user sequentially scanning an ordered list of document summaries. Residuals can be computed for all of these metrics by taking $r_i = 1$ for $i \notin J$, and these similarly provide an upper bound on the uncertainty in the measured score.

In related work, Robertson [25] proposes that the geometric mean of per-topic scores be preferred to the arithmetic mean, and suggests the use of GM-AP as an aggregate of AP scores, see also Ravana and Moffat [24]. The use of the geometric mean is not specifically tied to AP, and it can be applied to the per-topic scores generated by any other metric too. Indeed, the same over-all-topics system ordering attained by a "GM-M" variant of a metric $\text{M}(\mathbf{r})$ can be generated by defining $\log(\text{M}(\mathbf{r}))$ to be the "metric", and then aggregating in the usual manner by computing the arithmetic mean over those transformed values, including applying a t-test if so desired.

2.2 Comparing effectiveness metrics

Needless to say, effectiveness metrics behave differently in terms of the pairwise system relativities they induce, and hence also in terms of the multi-system orderings that they generate. *Shallow* metrics are strongly top-focused, and place substantial emphasis on relevance values near the head of the ranking. *Deep* metrics place less emphasis at the head of the ranking, so as to be able to spread influence further down the ranking. That means that when large numbers of documents are relevant, recall-based metrics are automatically "deep"; on the other hand, for topics that have smaller pools of relevant documents, recall-based metrics provide shallower evaluations. Compared to this, RR is a shallow metric regardless of the number of relevant documents, and, as an extreme example, precision at depth 1000 is always a very deep metric.

One of the features of RBP is that the choice of the parameter ϕ gives rise to different effective depths to the evaluation, varying from very shallow to very deep. In particular, the expected viewing depth in the RBP user model is given by $1/(1-\phi)$, which is 2 when $\phi = 0.5$, is 10 when $\phi = 0.9$, and is 100 when $\phi = 0.99$. Hence, for navigational web search a small value of ϕ might be appropriate, matching the user's expectations from the search and likely behavior during the search. In other applications, for example when a large pool of relevant documents is required in response to an informational query, a high value of ϕ is likely to be more suitable. In the TREC-8 Ad-Hoc Track relevance judgments that were mentioned earlier, the topics exhibit exactly this type of diversity, ranging between 6 relevant documents and 347 relevant documents, with a median of 70.5 and

a mean of 94.6. For the 1998 TREC-7 experiments [40], also involving 50 topics, the range was similarly 7 to 361 relevant documents per topic, with a median of 60 and a mean of 93.5.

When only one system's rankings are available, there is no reason – nor any sensible way of doing it – to compare metrics on a per-topic or averaged basis, since it is clear that the scores that the metrics give are incomparable. For example, there need not be any connection between the AP score for a particular run for a particular topic and the RBP score using some value of ϕ; both are alternative functions that take the particular ranked list and the relevance value associated with each item in the list as an inputs, and produce a single numerical score as an output. The difference in scores arises due to the set of underlying assumptions of how these fundamental inputs should be considered to give an understanding of search effectiveness, and the choices that are made to operationalize those assumptions.

2.3 Comparing system scores across topic sets

In a typical information retrieval experiment, the key comparison of interest is the relative effectiveness of a *pair* of systems where one is considered to be a *baseline* system, and the other is an *experimental* system that incorporates some change to the retrieval process. For a chosen test collection, and a chosen effectiveness metric (or sometimes a set of metrics), scores are first calculated for each topic. These individual scores are then aggregated, usually using the arithmetic mean, into a single overall effectiveness score for each system, and the system that achieves the higher score can be viewed as being "better" than the other system.

The two mean effectiveness scores are often analyzed further using a statistical significance test, to give the researcher confidence that the observed differences are not due to chance alone. A range of statistical tests have been used in IR research, and there has been ongoing debate about which are the most suitable [35]. In a recent systematic review of IR literature appearing in ACM SIGIR and TOIS from 2006–2015, Sakai [31] reported that the paired t-test is by far the most widely used procedure (66% and 61% of papers in SIGIR and TOIS that used a statistical test) followed by the Wilcoxon signed rank test (20% and 23%, respectively).

The paired t-test is used to evaluate the null hypothesis that two dependent samples represent two populations with the same mean values [34]; rejecting the null hypothesis therefore gives confidence that the two samples are likely to be from populations with different mean values. In test collection-based IR experiments, a paired test is typically appropriate since the same set of search topics is evaluated using both systems that are to be compared. The test procedure results in a p-value which gives the probability of observing the obtained result, or something more extreme, under the null hypothesis. This value can then be compared to a pre-determined significance level denoted here as α, and if the p-value is less than or equal to α, the outcome is deemed to be statistically significant [31]. Overall, the smaller the p-value, the higher the confidence that the measured difference is real rather than occurring by chance.

When two systems are being compared, a single p-value results, and lends (or not) credibility to the outcome of that particular system-versus-system comparison. On the other hand, shared-task experimentation in information retrieval – in which dozens of participating groups submit runs that are all scored using a common framework – creates the possibility of a large number of concurrent system-versus-system comparisons. For example, in one of the experimental contexts described in more detail in Section 3.1, a total of 65 systems each have scores over a set of 50 topics. In total there are thus $65(65-1)/2 = 2080$ pairs of systems. When operating at this scale, it is then possible to use the distribution of the set of resulting p-values as an indication of the usefulness of the measurement technique in question. For example, one value that can be derived is the *discriminative power* or *discrimination ratio* [27, 28], the fraction of the system pairs for which the corresponding p-value is less than some typical threshold such as $\alpha = 0.05$. The discrimination ratio can then be

thought of as being the odds of obtaining a statistically significant outcome from an experiment in which two of those contributing systems are chosen at random and then compared.

Each system-versus-system significance test is carried out based on per-topic performance scores derived using some particular effectiveness metric. If the same pair of systems is re-evaluated using a different metric, a different p-value will be computed, and the gross outcome of the significance test may also be different. That is, while metrics can be compared based on their discriminative power, it is also important to note that whether the metric in question reflects an attribute of the experimentation that it is valuable to measure is a quite different question and not one that we consider here; see, for example, the discussion provided by Moffat and Zobel [23] and by Fuhr [11].

2.4 Comparing system rankings

Pooled judgments usually arise as a consequence of experimentation in which a suite of systems are implicitly or explicitly being compared. The judgments are used to compute per-topic per-system metric scores, and then those scores are averaged across topics to obtain system average scores. Finally, those mean system scores can be used to order the systems from "best" to "worst". If some metric M_1 gives one ordering of the systems, and a second metric M_2 gives rise to another ordering, it is then natural to ask how alike or different M_1 and M_2 are in terms of the system orderings that they induce.

We employ two different methods for comparing pairs of system orderings. The first approach is to compute the well-known Kendall's τ coefficient. Each matched pair of items in the two n-element lists is either *concordant*, and appears in the same relative order in both lists, or is *discordant*, and appears reversed. Kendall's τ subtracts the number of discordant pairs from the number of concordant pairs, and then normalizes by $n(n-1)/2$, to obtain a value between -1 (one list is the reverse of the other) and $+1$ (the two lists have the elements in exactly the same order). Kendall's τ treats all pairs identically, and places as much emphasis on disorder at the bottom of the lists as it does at the top.

The second correlation we compute is the *rank-biased overlap* (RBO) between the lists [44]. Like its companion RBP, RBO is a top-weighted measure, and depending on the exact value used for its parameter ϕ, places increased emphasis on swaps that occur near the top of the lists. In addition, RBO has a range of other properties [44]. Because it is an overlap measure, RBO is zero when the two lists are disjoint, and 1.0 when they are identical. In terms of interpretation, the parameter ϕ is again a persistence adjustment, and RBO computes the expected fraction of items observed to appear in both lists by a randomized user when their probability of examining (only) prefixes of length x in the lists is given by $\Pr(x = d) = (1-\phi)\phi^{d-1}$.

We note that various alternative top-weighted correlation measures have been proposed, including the *AP correlation* (τ_{AP}), which is based on a probabilistic interpretation of the Average Precision effectiveness metric [46]. However, in a recent empirical analysis of the factors that influence the correlation between evaluation metrics, Ferro [10] concludes that while τ_{AP} and Kendall's τ might lead to different absolute correlation values for system rankings, both lead to consistent assessments in this context; we therefore report Kendall's τ in our experimental results.

2.5 Reliability of pooling and effectiveness measurement

Observing that it is not feasible to obtain exhaustive human relevance judgments for a query over a large collection of documents, Spärck Jones and Van Rijsbergen [36] proposed a technique called *pooling* whereby independent searches should be carried out to obtain more broadly based relevance judgments than would be available for a single system. In evaluation campaigns such as TREC, it is usual for a set of participating systems to be considered, and the union of their returned

documents to be judged; where this number exceeds the available budget for judging, the depth to which the contributing systems can recommend documents is constrained to a fixed rank [42]. The pooling process ensures that all systems that contribute to the pool are treated consistently, since they all have an equal opportunity to contribute to the set of documents that will be judged. However, when the same test collection is used to evaluate a new system that did not contribute to the pool, it is likely that some number of previously unjudged documents will be returned. The conservative default approach in IR experimentation is to treat any unjudged documents as if they are not relevant. However, this introduces a potential bias against new systems. There has therefore been extensive investigation into the reusability of test collections.

An analysis of the early TREC collections was carried out by Zobel [47] through "leave one out" experiments, where a system was re-evaluated after first removing any documents that were uniquely contributed to the pool by that system, effectively making them unjudged. The analysis led to the conclusion that existing test collections can be used to fairly evaluate new systems, while cautioning that the absolute performance of a system may be underestimated, and warning that researchers should consider the number of unjudged documents that new systems return.

Buckley and Voorhees [5] investigated the impact of incomplete judgments on the newer TREC-8, TREC-10 and TREC-12 test collections by progressively removing a randomly selected percentage of the full available qrels. The analysis demonstrated that as judgments become less complete, the Kendall's τ correlations between system orderings obtained using the full and reduced judgment sets begin to deteriorate. Buckley and Voorhees also proposed a new metric, BPref, and demonstrated that it retains a higher correlation than other metrics such as AP as relevance judgments are removed.

Sakai [28] proposed the use of *condensed* versions of the standard Q-Measure, AP and NDCG metrics, where unjudged documents are removed from the ranked list before the metric scores are calculated, and demonstrated that these are more effective than BPref in terms of both Kendall's τ rank correlation and discriminative power. This analysis was extended by Sakai and Kando [32] to further TREC and NTCIR collections, showing that the condensed metrics should be preferred over their non-condensed original formulations where unjudged items are present in ranked lists, and again concluded that they outperform BPref, as well as RBP, in terms of discriminative power and correlation with full-judgment system rankings. Like the earlier work of Buckley and Voorhees [5], Sakai and Kando employ relevance judgments that are randomly reduced to as little as 10% of their original size, measuring the effect of the reduction on system score correlation and metric discrimination power. Sakai [29, 30] explores the use of random judgment reduction versus depth-based reductions, including analysis of two types of bias that may be introduced by random reduction.

In addition to the issue of unjudged documents when re-using a test collection, other factors that may potentially bias results have also been studied. As test collections continued to increase in size, concerns about the pool being a representative sample arose again, with Buckley et al. [4] reporting a favoring of documents that contain query terms in the title, presenting a risk that new systems that use wholly different retrieval approaches may not be evaluated fairly with such a test collection.

A further factor that may impact on the reliability of test collection-based evaluation of IR systems is the relevance judgments themselves. Relevance is a complex concept that may include static aspects such as the topical content of a document, and dynamic aspects such as novelty [18]. To avoid these complexities, relevance judgments for test collections typically focus on the topical "aboutness" relation between a query and a document, and require that each document be judged independently on its own merits. Despite this, human assessors may still disagree on whether particular documents are in fact relevant for an information need. Voorhees [37] investigated the

impact of such disagreements on system effectiveness evaluations by comparing the extent to which system orderings are affected when using different sets of relevance judgments made by expert assessors and by students. The analysis demonstrated that although individual relevance judgments may vary, system rankings are robust to such differences, with a Kendall's τ correlation of around 0.9. Attributes of the human assessors – in particular, whether they authored a search topic statement or not, and their level of knowledge about the search topic – have also been found to impact on the consistency of test collections [2].

We note that a range of alternative approaches for the construction of test collections have been proposed in the literature, including: minimal test collections, where documents are selected for judging based on their ability to discriminate systems [7, 8, 22]; using machine learning classification techniques trained on judged documents to predict the relevance of unjudged documents [6]; and using online learning techniques to determine which documents are most likely to be relevant and should therefore be judged [1]. Recent work by Losada et al. [15] and Lipani et al. [13, 14] has continued to explore non-uniform pooling strategies. However, uniform fixed-depth pooling remains in wide use, and was the basis for the construction of the TREC collections that we use here. Reflecting that pattern, in our experiments in this work we use fixed-depth pooling only, with uniform selection to a range of depths d' as a way to deterministically create reduced judgment sets, rather than via random removal. Other recent work has considered the necessity for pooling not just over systems, but also over a set of query variations for each of the topics [3, 20].

There has also been work undertaken in terms of metric evaluation depth, as distinct from judgment pooling depth [43]. For utility-based metrics such as RBP, relative system scores tend to be stable as the evaluation is deepened, because scores are non-decreasing. But for truncated recall-based metrics, extended evaluation beyond the pooling depth may lead to substantial changes in system ordering [16].

Given this extensive range of prior work as context, we arrive at a critical question that we consider in this paper: with utility-based metrics, the residual provides an (albeit, pessimistic) upper limit on the extent of measurement uncertainty in a computed system score. Is there any such equivalent that can be inferred in connection with recall-based metrics such as AP and NDCG? And, if there is, does it indicate anything in regard to the usefulness of a paired system-versus-system comparison over a set of topics?

3 MEASUREMENT OF RELIABILITY

3.1 Datasets and methodology

We make extensive use of resources that have been compiled by the NIST-sponsored TREC initiative, see Voorhees and Harman [42] for details of this long-running endeavor. In each round of TREC experimentation a set of system runs were developed by research groups at universities and commercial organizations, using a defined collection of documents and a set of 50 or more topic statements, and submitted for evaluation. Some subset of those runs were next pooled to create judgments, with the subset typically selected so as to ensure that all of the research groups that had created the runs had approximately the same number of runs contributing to the pool. Those judgment were then used to compute effectiveness scores for each system for each topic, and then aggregate (usually arithmetic mean) scores for each system. A number of metrics were used in connection with each round of experimentation, notably including average precision (AP) in all three of the newswire collections we use here: the TREC-7 Ad-Hoc Track and topics 351–400; the

Dimension	Collection		
	TREC-7 Ad-Hoc	TREC-8 Ad-Hoc	TREC-13 Robust
Year	1998	1999	2004
Number of topics	50	50	249
Number of systems	103	129	110
Number of systems pooled	77	71	*n/a*
Pooling depth	100	100	*n/a*
Number of documents judged (avg.)	1606.9	1736.6	1250.6
Number of relevant documents (avg.)	93.5	94.6	74.1
Depth of first unjudged document (avg.)	101.8	95.8	69.6
Number of deeply-judged systems	65	67	0

Table 1. TREC collections and qrels used in experimentation. The values for documents judged and relevant documents are per-topic averages. The second to last row gives the average (across systems and topics) rank at which the first unjudged document appears in each run. The last row gives the number of systems that generated a run of at least 50 documents for every topic *and* had every document judged down to a rank of at least depth 50.

TREC-8 Ad-Hoc Track and topics 401–450; and the TREC-13 Robust Track and topics 301–450, plus topics 601–700, excluding topic 672.[1]

Table 1 lists a range of parameters for each of those three different TREC experimentation rounds, including the number of topics, the total number of systems, the number of those that were pooled, the average number of documents judged per topic, and the average number of those judgments in which the document was determined to be relevant. Note that in each experimentation cycle around 5–6% of documents judged were deemed to be relevant. Note also that the 2004 TREC13 Robust Track judgments were an amalgam of fresh judgments that year and judgments compiled in several previous years [38, 39], and is why two of the entries are marked as "n/a".

In all of these three experimental rounds, the per-system final reports provide effectiveness scores in terms of recall, precision at a range of depths, R-Prec, and AP. The latter, aggregated across topics via the arithmetic mean, is probably regarded as being the dominant assessment. In the 2004 TREC-13 Robust Track, an adjusted GM-AP metric [25] was also considered, with $\epsilon = 0.00001$ added to each raw score before the averaging process, and then subtracted again from the computed average [38].

Noting the observations of Yang et al. [45], we re-sorted all of the submitted runs associated with these three tracks, so that the correct ordering was generated even when documents scores were represented using exponential notation (column five in each run). Ties on score were decided according to the assigned rank at the time the run was constructed (column four in the submitted file) with a sort by document identifier the final step to ensure that the ordering was deterministic.[2] This ensures that documents are considered in order of decreasing score as specified by the implementors of each system. Re-sorting resulted in different orderings for some subset of the topics for the great majority of systems, for all three collections. The numbers presented in Table 1 and in the remainder of this work are in all cases with respect to the re-sorted runs, and all further references to run and rank cut-offs within runs are based strictly on the resultant ordering, with no further

[1] There were no relevant documents identified by the pooling process for this topic, which means that recall-based effectiveness metrics cannot be computed.

[2] `sort -k1,1n -k5,5gr -k4,4n -k3,3`, with the final "`-k3,3`" component not a deciding factor in any of the runs.

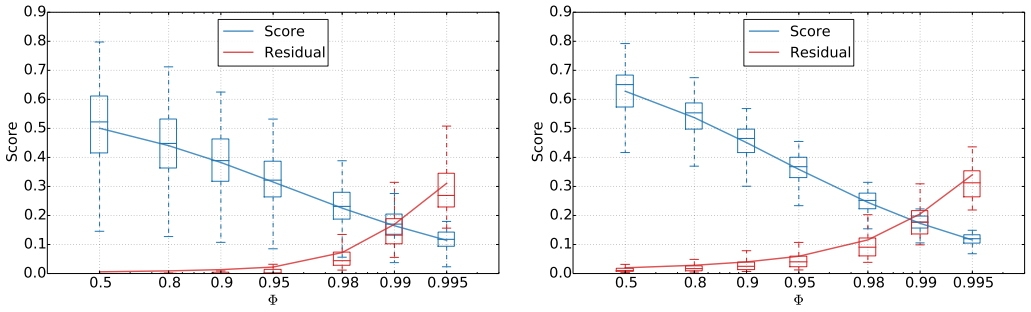


Fig. 1. TREC-7 (left) and TREC-13 Robust (right), distribution of RBP scores and residuals as a function of ϕ. Each box/whisker element reflects the set of scores attained over all systems and all topics for that value of ϕ. The horizontal scale is determined by the logarithm of the expected depth, $1/(1-\phi)$, with the labeled values of ϕ corresponding respectively to expected depths of 2, 5, 10, 20, 50, 100, and 200 documents.

attention paid to the scores and ranks embedded in the runs provided by the participating research groups.

The last row of Table 1 reports the number of *deeply-judged systems*, defined as systems for which every topic in the test set gave rise to a run containing at least fifty documents, and where as a minimum every document in the first fifty was judged for every topic. The discrepancy between this value and the nominal number of systems pooled as reported in the track overviews [38, 40, 41] arises because some of the pooled systems generated a short run of fewer than fifty documents for at least one of the topics. The zero value reported in this dimension for the TREC-13 collection is a consequence of the multi-year process used to create the qrels file – clearly it was not possible for the TREC-13 systems to contribute to the prior-year pools that generated the judgments for the 200 carry-over topics. When restricted to the 49 new topics created, pooled, and judged in 2004, there are 52 deeply-judged systems (and a total of 42 systems that contributed to the pool).

Where a run of length k had every document in it judged (including in the case of short runs), the document at rank $k+1$ was deemed to be the first one unjudged for the purposes of computing the average depth of the first unjudged document values, shown in the second-to-last row.

3.2 Behavior of RBP

Figure 1 shows typical patterns of RBP scores and residuals for two TREC collections. Each plotted element represents the range of RBP scores (blue) and RBP residuals (red) over all systems and all topics. When ϕ is small, the evaluation is shallow and focused on a relatively small number of documents at the top of the rankings, and hence the residuals are small. On the other hand, measured RBP scores are also quite high, because on average the systems are able to bring relevant documents into the top few positions in the ranking.

However, as the parameter ϕ increases, the extent of uncertainty in the measurements also increases, because a smaller fraction of the assessment weight is near the top of each ranking, meaning that a smaller fraction of the documents involved in the assessment are pooled and hence judged. At the same time, the RBP scores decrease, partly because of the unavailability of all of the needed judgments, and partly because the systems are not as good at placing relevant documents into position (say) 50 as they are into position one. Unsurprisingly, in all three of the collections (the TREC-8 graph is similar), the residual exceeds the measured RBP score once $\phi \approx 0.99$ and the expected depth of the evaluation is approximately 100, the pooling depth.

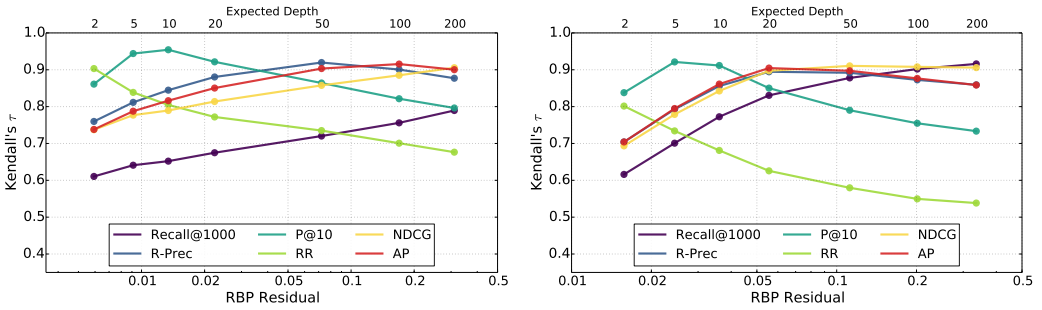


Fig. 2. TREC-7 (left) and TREC-13 Robust (right), Kendall's τ correlation between the final system orderings induced by RBP and a range of ϕ parameters, and six other metrics, plotted as a function of the average RBP residual (see Figure 1). Each plotted point corresponds to one value of ϕ, from 0.5 on the left through to 0.995 on the right, and with the corresponding expected evaluation depths varying from 2 to 200, respectively, noted informally across the top of each of the two graphs. The lower scale showing the resulting residual is correlated with but not an exact function of that depth; and the relationship also differs between the two rounds of experimentation.

Note that there is no sense in which the RBP residual should be thought of as being a "confidence interval". Rather, it is an optimistic estimate of how much the measured score could increase were all unjudged documents to be relevant. A more measured estimate, but perhaps still a relatively generous one, would be to suppose that if the unjudged documents were to be judged, they would be found relevant at roughly the same 5–6% rate as the documents in the pooled set. If this were the case, then a reasonable supposition might be that the "true" RBP score was similarly 2.5–3% larger than the measured score when $\phi = 0.99$, since that is (broadly speaking) the cross-over point at which score and residual are equal. Lu et al. [17] explore mechanisms for generating estimates of relevance for unjudged documents, based on their positions in runs and the relevance labels of the judged documents in the same runs.

3.3 Estimating ϕ for other metrics

Given that the RBP residual can be bounded above if both the parameter ϕ and the pooling depth are known [23], a natural question is to ask whether there are particular values of the RBP ϕ parameter that correspond closely to other metrics – in particular, to recall-based ones.

The two panes in Figure 2 plot values of Kendall's τ, comparing the overall system orderings created by a range of standard reference metrics with the orderings generated when RBP is used to score the systems, across a range of RBP parameters. The horizontal axis reflects the mean RBP residual across systems and topics, with each marked point on each curve corresponding to one of the values of ϕ plotted in Figure 1. The corresponding expected viewing depths are listed across the top of the graph. The region of highest Kendall's τ for each of the plotted curves shows the range of ϕ for which RBP yields a system ordering closest to that generated by the corresponding reference metric. For example, RBP is most like Prec@10 when $\phi \approx 0.9$ and the expected viewing depth in the RBP user model is 10. The four recall-based metrics that are plotted – R-Prec, Recall@1000, AP@1000, and NDCG@1000 – all have deeper evaluation patterns, and have their maximum correlation with RBP when ϕ is higher, with several of them peaking when $\phi = 0.99$ (with the same pattern evident in a third plot for TREC-8, not included here).

The relationships shown in Figure 2 are not unexpected. That RR and Prec@10 are shallow effectiveness metrics, and that AP and NDCG are deep effectiveness metrics, is well understood.

It is also known that high values of ϕ correspond to deep evaluation [23]. Nevertheless, Figure 2 provides evidence to suggest how deep AP, NDCG, and R-Prec actually are; and more importantly, shows that "AP-like", "NDCG-like", and "R-Prec-like" evaluations correspond to (typically) ϕ values of (looking at the upper axis) 0.98 or more, and hence RBP-based residuals of (looking at the lower axis) 0.1 or more.

Voorhees [37] analyzed the effect that using different human relevance judgments can have on system ranking correlations. Comparing judgments from TREC assessors versus university students, the results showed Kendall's τ correlations in the range from 0.87 to 0.95 between rank orderings of systems that participated in TREC-6 for different combinations of judgments. This further suggests that the level of correlation observed in our data is high, and that any remaining discrepancies are likely to be no greater than what might be observed by taking into account human variation in relevance assessments.

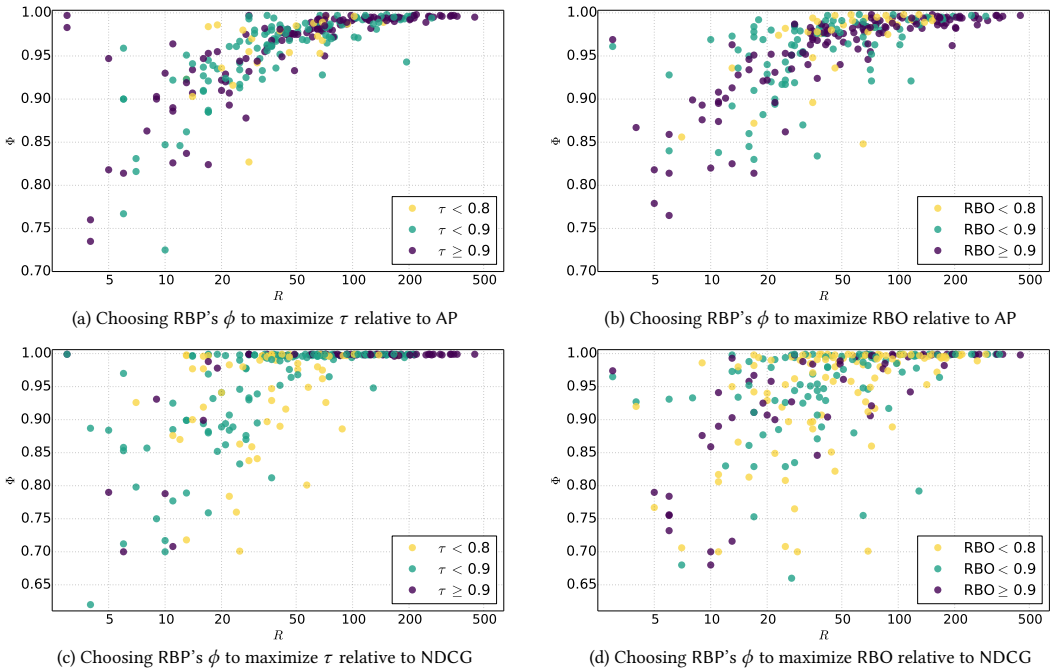
3.4 RBP parameter variation as a function of R

In Figure 2 the system orderings used to compute the correlations were based on mean scores, computed by averaging each of two metrics across the same set of topics. That led to a single τ correlation score as each pair of metrics was compared, where RBP becomes a "different" metric each time ϕ is changed. But it is also possible to generate a separate system ordering for each topic in the test collection, and compute per-topic correlation coefficients.

Figure 3 shows the results of carrying out such an experiment. To generate each of the four scatter plots, a reference metric was selected, AP@1000 or NDCG@1000, and then the topics were processed one by one. For each topic, all of the systems were scored using the chosen metric and the relevance judgments, and a system ranking generated based on those single-topic scores. The same systems were then scored for that topic using RBP, with a search over the ϕ parameter space carried out from $\phi = 0.500$ to $\phi = 0.999$ in 0.001 increments. Across the values of ϕ, the one that gave the system ordering with the greatest correlation score relative to the ordering of the reference metric was noted, together with the correlation coefficient. In the case of ties of coefficient, the smallest maximizing ϕ was the one that was used. The set of maximizing ϕ values (vertical axis) was then plotted as a function of the number of relevant documents for that topic (horizontal axis), with the strength of each individual correlation indicated by the color of the plotted dot. The four panes in the figure cover two metrics, AP and NDCG, and two correlation coefficients, Kendall's τ and RBO using $\phi = 0.9$.

The clear pattern that emerges from the top two graphs, based on seeking to "fit" against AP, is that when R, the number of known relevant documents for that topic, is small, then the "most similar" RBP-based ordering is also achieved when ϕ is relatively small. Conversely, when R is large for a topic, then the RBP-based system ordering is closest to that of AP when ϕ is large. A similar outcome results when NDCG is used as the reference metric (the two lower panes), with, for the most part, small values of R best fitted by choosing small values of the RBP parameter ϕ. Table 2 summarizes the plotted relationships, giving Kendall's τ correlation coefficients and significance values for the four sets of points plotted in Figure 3 (the τ of the τ's); together with a summary of the correlation score distribution, in effect counting the number of plotted points of each color in each of the four graphs. Table 2 also lists the average (over all 249 topics) value of ϕ for each of the four situations reported.

Note that these outcomes are not intended to be construed as an argument that RBP should be used with a value of ϕ that is determined on a topic by topic basis as a function of R. That would then suggest – as is the case with all recall-based metrics – a user model in which the user was aware of R prior to having seen any of the ranking, which is unrealistic. Rather, the user's primary influence in determining their behavior is the total volume of relevance they seek to identify, the



(a) Choosing RBP's ϕ to maximize τ relative to AP

(b) Choosing RBP's ϕ to maximize RBO relative to AP

(c) Choosing RBP's ϕ to maximize τ relative to NDCG

(d) Choosing RBP's ϕ to maximize RBO relative to NDCG

Fig. 3. TREC-13 Robust, relationship between R, the number of known relevant documents (horizontal axis) for a topic, and the per-topic value of ϕ that maximizes a rank-correlation coefficient computed for the system orderings induced for that topic, for two different correlation coefficients and two different recall-based metrics. In the top row, the reference metric is AP in both panes; in the second row it is NDCG. Kendall's τ is used as the correlation coefficient in the left column, and RBO (computed using $\phi = 0.9$) is used in the right column. One point is plotted for each of the 249 topics in each of the four panes.

Metric	Correlation	Count of cases			Kendall's τ	p	Average ϕ
		< 0.8	< 0.9	≥ 0.9			
AP	Kendall's τ	23	110	116	0.66	< 0.0001	0.958
AP	RBO, $\phi = 0.9$	36	90	123	0.59	< 0.0001	0.953
NDCG	Kendall's τ	58	131	60	0.51	< 0.0001	0.956
NDCG	RBO, $\phi = 0.9$	99	95	55	0.46	< 0.0001	0.943

Table 2. Strength of correlations between two recall-based reference metrics and RBP when ϕ is chosen to maximize the relationship on a per-topic basis, using the 249 topics of the TREC-13 Robust collection. The final column shows the average (over topics) value of maximizing ϕ for that configuration of metric and correlation measure.

target T proposed by Moffat et al. [20, 21]. Our purpose in this section has been to show that if we wish to closely match the behavior of recall-based metrics with utility-based ones so as to be able to estimate residual-like error limits for the recall-based metrics, we should do so based on a knowledge of R.

Dimension	Collection		
	TREC-7 Ad-Hoc	TREC-8 Ad-Hoc	TREC-13 Robust
Topics	50	50	49
Systems	65	67	52
Documents, judged	33,870	40,238	17,509
Documents, relevant	3121	3175	1576
Single-vote documents, judged	15,943	24,149	5597
Single-vote documents, relevant	639	731	147

Table 3. Reduced qrels files when the deeply-judged runs (see Table 1) and their top-$d' = 50$ ranks are pooled. The last four rows provide aggregates across all topics and all systems, covering, respectively: the number of distinct topic-document pairs in the reduced pool; the number of those documents that were relevant according to the NIST qrels; the number of those that were members of the pool as a result of being nominated by a single system; and the number of those "single nomination" documents that were judged relevant. The TREC-13 columns refer to topics 651–700 (minus topic 672) only.

3.5 Adding uncertainty via reduced qrels

The next experiment adds imprecision to each system-topic score, by supposing that only a subset of the judgment pool is available when evaluating each topic. We are interested in exploring the connection between residual – and by implication, the fidelity of the measured score – and the ability of the measurement regime to separate systems. One way of quantifying the latter is via a statistical test, and the p-value that is then generated.

To create reduced judgment sets that equally disadvantage all systems, we start with the set of deeply-judged systems to depth of at least $d = 50$ (see Table 1), and apply a set of artificial pooled judgment depths of $d' = \{5, 10, 15, 20, 25\}$, and $d' = 50$. For example, in the case of the TREC-7 collection, the runs of the 65 deeply-judged systems, across 50 topics, were then top-5 filtered, top-10 filtered, and so on through until top-50 filtered, and in each of the six cases, those selected documents' entries (and only those entries) were extracted from the NIST qrels file to make a reduced qrels file. The same procedure was also applied to the TREC-8 and TREC-13 submitted system runs and qrels. The result is a set of qrels files in which all pooled systems were given demonstrably equal opportunity to provide documents and have them judged.

Table 3 provides information in connection with the $d' = 50$ qrels files. For example, in the case of TREC-7, the top-50 for each of the 65×50 system-topic combinations ($65 \times 50 \times 50 = 162{,}500$ documents in total) resulted in a reduced set of 33,870 unique documents that retained their labels into the reduced qrels file; of those, 3121 documents had been previously judged to be relevant. Each qrels file was then further processed to identify the number of systems that had nominated each of the documents. The number of documents with only a single nomination (over the set of deeply-judged systems) is also shown in Table 3, along with the number of that subset that were judged relevant. Note that in the case of the TREC-13 Robust Track, this experiment was restricted to the matching set of judged topics, 651–700, not including topic 672. The smaller number of deeply-judged systems for this collection means that there is a correspondingly smaller number of documents judged per topic.

Table 4 decomposes a different reduced qrels file, with $d' = 20$, according to the number of different systems that nominated each document; and calculates conditional probabilities of relevance as observed in the reduced qrels file. There is a clear pattern here that the greater the number of systems that had any particular document in their top $d' = 20$, the greater the chance of that document being deemed relevant by the NIST assessors. If only a single system nominates a

Multiplicity	TREC-7 Ad-Hoc		TREC-8 Ad-Hoc		TREC-13 Robust	
	Count	% Rel.	Count	% Rel.	Count	% Rel.
1	6988	7.8	10,266	5.3	2329	4.7
2	2791	9.5	2538	11.7	1147	8.0
3–4	1651	17.4	1548	16.3	1132	9.2
5–8	1178	23.1	1017	26.6	1010	14.2
9–16	780	29.5	732	34.0	654	25.5
17–32	567	44.3	544	43.6	552	38.9
33+	375	65.1	416	61.3	369	61.8
Total	14,330	14.6	17,061	12.3	7193	14.7

Table 4. Probability of a document being judged relevant, as a function of the number of systems that nominated it (the document's *multiplicity*) into a pool formed to a depth of $d' = 20$. Documents that appear in the top-20 of more than half the pooled systems (the 33+ band for TREC-7 and TREC-8) have a higher than 60% chance of being relevant. The TREC-13 columns refer to topics 651–700 only.

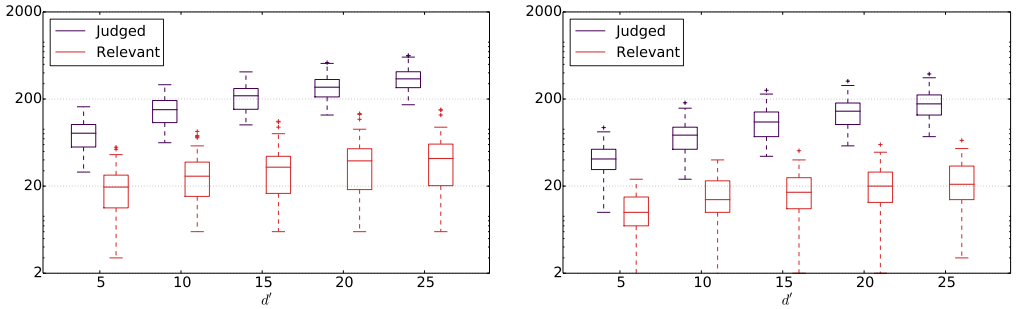


Fig. 4. TREC-7 (left) and TREC-13 Robust (right, topics 651–700 minus 672), the number of documents and the number of relevant documents in the reduced pools, both measured on a per-topic basis, for depths $d' = 5, 10, 15, 20$ and 25. Note the logarithmic vertical scale.

document, the observed probability of being relevant is under 10%, but if 33 or more of the pooled systems include that document in their top-20, that probability is over 60%. Similar data for $d' = 10$ shows even higher conditional probabilities, while for $d' = 30$, $d' = 40$ and $d' = 50$ the probabilities are lower compared to those for $d' = 20$.

Figure 4 shows the range of pools sizes across topics, as the pooling depth d' varies from 5 to 25. Each pair of box/whisker elements reflects the distribution of pool sizes at that pooling depth, and the corresponding distribution of relevant documents, as identified by runs associated with the deep-judged systems. As the pool depth d' increases, so too does the number of documents in the pool. The number of relevant documents identified as the pool is extended also increases, but at a slower rate, and the declining rate of discovery can be used as a basis for estimating R, the total number of relevant documents for each topic [47].

Figure 5 shows how depth of judgments affects metric score for three different metrics, as d' varies from 5 to 25. Rank-biased precision scores – and any other weighted-precision approach – of necessity are non-decreasing as the judgment pools are extended. That is, the scores obtained through the use of any particular qrels file cannot be greater than the scores obtained after further judgments are added. But AP and NDCG scores typically (but not monotonically) *decrease* as judgment pools are deepened. This divergent behavior occurs because recall-based metrics include

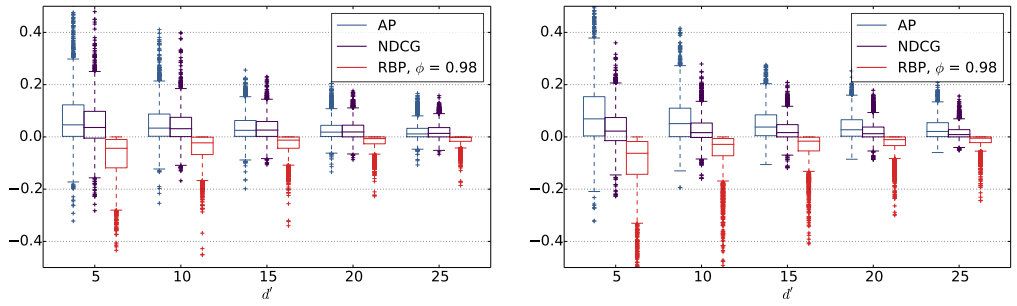


Fig. 5. TREC-7 (left) and TREC-13 Robust (right), evolving score differences across combinations of systems and topics for AP, NDCG, and RBP ($\phi = 0.98$), for pooling depths $d' = 5, 10, 15, 20$ and 25, in each case relative to a reference point established by the corresponding $d' = 50$ score for that system-topic combination. Only the deeply-judged systems are used.

Collection	Metric	d'					
		5	10	15	20	25	50
TREC-7	AP	68.0	70.9	71.7	73.2	73.3	73.0
	NDCG	71.2	73.5	74.1	74.4	74.8	75.0
	RBP, $\phi = 0.98$	68.4	69.7	70.6	70.6	70.4	69.5
TREC-8	AP	71.1	72.6	73.0	73.6	73.5	73.8
	NDCG	68.7	70.2	70.2	70.8	71.3	71.6
	RBP, $\phi = 0.98$	69.6	71.2	71.7	72.1	72.2	72.5
TREC-13	AP	63.4	63.1	63.7	64.5	66.1	66.7
	NDCG	58.1	59.3	61.1	60.9	62.1	64.7
	RBP, $\phi = 0.98$	57.0	55.5	56.0	54.3	56.0	56.9

Table 5. Measured discrimination ratios: the percentage of deeply-judged system pairs for which a significant difference is indicated, as d' varies from 5 to 25, and for $d' = 50$. In all cases p is the result of a paired two-tailed t-test, thresholded at $\alpha = 0.05$.

a normalization by R, the number of relevant documents that have been identified for that topic (or a function of it in the case of NDCG), and R is non-decreasing as pools are deepened. Moreover, the slowing rate at which relevant documents are encountered in any particular run as documents are considered at deeper depths means that growth in the "numerator" component of the recall-based metrics is insufficient to overcome their "denominator" factors, and computed scores decline.

Table 5 then shows the effect that pooling depth d' has on discrimination ratios. To compute these values, each possible pair of deeply-judged systems (for example, in the case of TREC-7, $65 \times 64/2 = 2080$ such pairs) was treated as being a "system comparison" over the topic set (in the case of TREC-7, 50 topics) using one of the reduced qrels sets for $d' \in \{5, 10, 15, 20, 25, 50\}$, with the Student t-test applied to the set of paired metric scores to compute a p-value. The fraction of those p-values less than $\alpha = 0.05$ was then counted. Note that there was a small number of instances where pairs of submitted systems from the same research group generated exactly the same set of scores, and p-values were not computable. The three system pairs in this category have been excluded from all of the results presented in this section.

d'	AP				NDCG				RBP, $\phi = 0.98$			
	TP	FP	FN	TN	TP	FP	FN	TN	TP	FP	FN	TN
5	1355	60	164	501	1428	52	133	467	1320	103	125	532
10	1416	58	103	503	1488	40	73	479	1371	78	74	557
15	1452	40	67	521	1510	32	51	487	1395	73	50	562
20	1484	39	35	522	1522	25	39	494	1416	53	29	582
25	1494	30	25	531	1534	21	27	498	1425	39	20	596

Table 6. Difference in significance test findings between evaluation using judgments to depth $d' = \{5, 10, 15, 20, 25\}$ and to depth $d = 50$, taking the latter to be the correct outcome. Each group of four values represents the count of true positives (TP, where the assessments to depth d' and to depth $d = 50$ both indicate significance); false positives (FP, where the shallow evaluation indicates significance, but the deeper $d = 50$ evaluation does not); false negatives (FN, where the deep evaluation indicates significance, but the shallow evaluation does not); and true negatives (TN, where neither evaluation indicates significance). All results are for the 2080 system pair combinations generated by the 65 deeply-judged TREC-7 systems, with significance measured using a two-tailed paired t-test and the threshold $\alpha = 0.05$.

As can be seen from Table 5, the two recall-based metrics typically have slightly higher discrimination ratios than does RBP, even when a relatively high value of ϕ is used. The relationship between AP and NDCG is less clear cut. What is perhaps surprising in Table 5 is that discrimination ratios do not uniformly increase with pooling depth d'. In most cases there is a small gain when deepening the pool from $d' = 5$ to $d' = 25$, but there are also exceptions, and only very small additional gains occur as the depth is further increased to $d' = 50$. That is, adding further evidence to a system-versus-system comparison (in the form of deeper judgments) does not necessarily lead to a greater degree of statistical confidence in the outcome.

3.6 Consistent discrimination

A key concern that then arises is the extent to which the set of system pairs that are found to be significant is the same at each pooling depth d'. Table 6 provides a detailed breakdown, using the TREC-7 collection and the 65 deeply-judged systems associated with it (see Table 1). For each metric and shallow pool depths $d' = \{5, 10, 15, 20, 25\}$, each of the 2080 system pairs was categorized as being one of TP (both the shallow depth d' judgments and the deep $d = 50$ judgments led to an indication of statistical significance); FP (the shallow judgments led to an indication of significance, but the deeper judgments did not); FN (the $d = 50$ judgments indicated significance, but the shallower depth d' judgments did not); and TN (neither set of judgments led to an indication of significance). Similar patterns of behavior (not shown here) arose in connection with the deeply judged systems in the TREC-8 and TREC-13 experimental rounds.

We also checked all of the true positive (TP) situations that arose to ensure that in both the depth d' and depth d evaluations the two system means (over the topic sets) had the same directional relationship. No situations in the TREC-7 systems and TREC-13 systems were detected where significance at both d' and $d = 50$ was identified but favoring different systems, but there were three such "contradictory" system pairs in the TREC-8 systems when evaluated using RBP and $d' = 5$. These three pairs were counted as false positives (FP) rather than as true positives (TP).

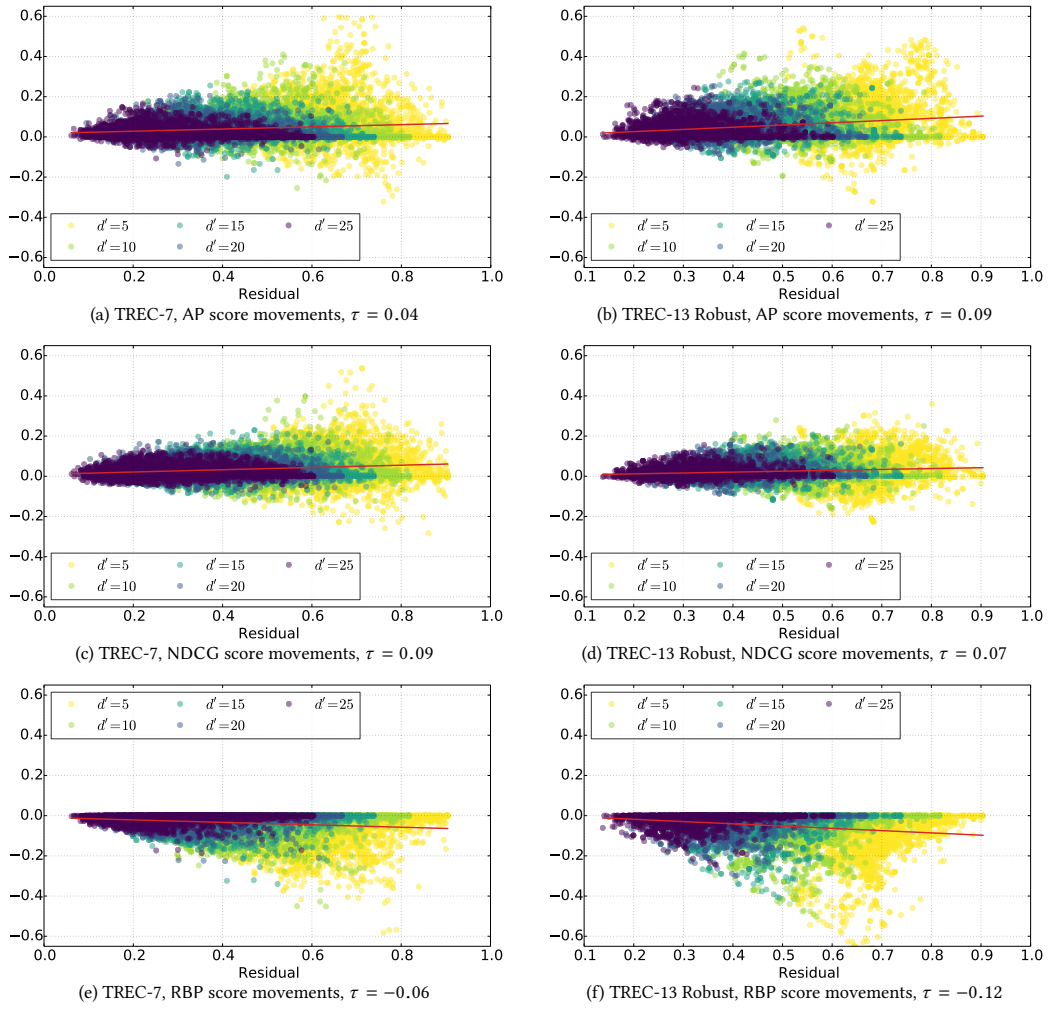
Within each group of four values the discrimination ratio is given by (TP+FP)/(TP+FP+FN+TN); what we focus on now is the FP column, which counts instances where the use of shallow judgments pooled to depth d' would lead to a conclusion of a significant difference between systems that would then be reversed by a more comprehensive evaluation based on pooling to depth $d = 50$. Looking

834 at the three FP columns, there is a uniform decrease in the false positive rate as the pool depth
835 d' is increased – not unexpected, since as d' increases the judgment set monotonically converges
836 towards the $d = 50$ judgment pool that is used to provide the reference point. Also worth noting
837 is that the number of FP instances is reassuringly small, even when $d' = 5$ and a very shallow
838 judgment pool is in operation. On the other hand, the monotone decrease in the size of the FP set
839 as d' increases does mask slightly more complex behavior. For example, for AP, the decrease from
840 60 to 58 as d' increases from 5 to 10 involves 13 system pairs moving out of the FP column, and a
841 further 11 system pairs shifting into the FP count. Further down the same column, the decrease
842 from 39 to 30 is the result of 12 system pairs leaving the FP set, and 3 system pairs entering it.
843 Given the nature of statistical testing and the abrupt binary cutoff introduced by the threshold
844 $\alpha = 0.05$ that is behind these counts, this level of drift is to be expected.
845
846
847
3.7 Back to residuals

848 Having observed that system-topic scores for recall-based metrics are affected in different ways
849 as a result of judgments being added to a pool, and that the strength of a system-versus-system
850 multi-topic comparison can shift in a quite unpredictable manner as judgment pools are deepened,
851 we return to RBP-based residuals, and examine one further question: the strength (if any) of the
852 relationship between the residual associated with a run, and the score movement that takes place
853 if more documents are judged. In this experiment, we consider each system-topic combination,
854 and take the corresponding $d' = 50$ score as a reference point, regarding it as being the "closest
855 to final" score that we can compute over the set of deeply-judged systems. For each of the other
856 $d' \in \{5, 10, 15, 20, 25\}$ values, we then compute the difference between the score at that depth and
857 the score at $d' = 50$ (the same differential that was plotted in overall terms in Figure 5), and plot a
858 point based on the computed RBP residual for that run (using $\phi = 0.98$). The resultant scatter-plots
859 – covering three metrics and two collections – are shown in Figure 6.

860 In all six of the panes, larger values of d' (as indicated by the more darkly colored points) lead to
861 both smaller RBP residuals and a correspondingly smaller range of score differences relative to the
862 $d' = 50$ evaluation. The two recall-based metrics AP and NDCG both behave in broadly the same
863 way, with the majority of the score differences positive, but a minority negative. That is, in most
864 cases, adding judgments will decrease the measured scores. In contrast, the weighted-precision
865 metric RBP (in the final two panes) of necessity has strictly non-positive score differences. Kendall's
866 τ correlation coefficients for the first four panes are all positive, but relatively small, and there is
867 only a weak relationship between residual and score difference in terms of expected outcomes. In
868 the case of RBP the τ values are less than zero, but again relatively small.

869 It is perhaps surprising that there is only a very weak correlation between RBP residual and
870 score changes for RBP, AP, or NDCG. A possible explanation for this is that the system-topic
871 combinations with high residuals are ones that were "unusual" in some way, and as a result had
872 selected a high fraction of documents that were unlabeled by other systems. This in turn indicates
873 a lower conditional probability of finding relevant documents (see Table 4), and hence that runs
874 with high residuals are also more likely to be low-scoring ones, for which large score differences
875 are unlikely to occur as the pool is extended. To test this effect, we also computed τ correlations
876 when, rather than arithmetic score difference, the second coordinate (y-axis) was the relative score
877 difference: the ratio of the metric score at depths $d' < 50$ to the corresponding score at $d' = 50$.
878 For the six conditions illustrated in Figure 6, all of the correlations became stronger: in order (a) to
879 (f), they were 0.18, 0.14, 0.16, 0.08, −0.19 and −0.23. (Note that in the case of TREC-7 a total of 35
880 system-topic instances were not used in this computation, because no relevant documents were
881 identified in the top $d' = 50$ ranks; and in the case of TREC-13, 14 system-topic instances were
882



(a) TREC-7, AP score movements, $\tau = 0.04$

(b) TREC-13 Robust, AP score movements, $\tau = 0.09$

(c) TREC-7, NDCG score movements, $\tau = 0.09$

(d) TREC-13 Robust, NDCG score movements, $\tau = 0.07$

(e) TREC-7, RBP score movements, $\tau = -0.06$

(f) TREC-13 Robust, RBP score movements, $\tau = -0.12$

Fig. 6. Score movements as a function of RBP residual for $d' \in \{5, 10, 15, 20, 25\}$, relative to judgments based on $d = 50$, computed on a per-system per-topic basis using $\phi = 0.98$. Three metrics (rows) and two collections (columns) are illustrated. Positive values on the vertical axis indicate computed metric scores for system-topic combinations at shallow depths $d' < 50$ that are greater than the corresponding $d = 50$ score. The colors indicate different values of d'.

removed.) That is, the correlations are all mildly stronger if score ratios are used rather than score differences.

To further investigate the relative effects caused by judgment pool variations, topic difficulty variations, and system variations, we also carried out an ANCOVA (analysis of covariance) analysis of the scores arising from the application of four different effectiveness metrics (AP, NDCG, RBP and RR). Using the 65 deeply-pooled retrieval systems selected from the TREC-7 experimental environment, the 50 corresponding topics, and a total of 10 different judgment pool depths as the covariate, we take as the null hypothesis that there is no variation in metric score caused by any of system, topic, pooling depth, or interactions between these. The shallow metric reciprocal rank

(RR) was included in this experiment as a reference point for which a smaller effect attributable to d' might be expected. However, the results of the analysis showed both significant primary effects and significant interactions ($p < 0.0001$) for all variables (pool depth, topic, and system); these outcomes held for each of the four effectiveness metrics, and across all three collections.

4 CONCLUSIONS

We have used a number of approaches to allow estimations to be made of the reliability of the scores developed by recall-based effectiveness metrics. The standard approach to estimate the reliability of a system-versus-system comparison is to use a paired t-test, on the assumption that any imprecision in the scores will show up as a higher p-value, and hence lower confidence in the outcome of the experiment. The results presented in Section 3 show that this presumption is not necessary reliable. In experiments in which uniform-depth pooling arrangements are systematically degraded, it is not the case that shallow judgment pools give rise to a loss of confidence in experimental outcomes. Indeed, discrimination ratios remain relatively unaffected even when metric scores are computed based on quite shallow pooling, and when the metric scores themselves are open to considerable imprecision.

We have also considered the residuals that are computed when a utility-based metric is used, focusing on RBP. We have demonstrated that the system orderings induced by recall-based metrics such as AP and NDCG can be quite closely approximated by RBP using relatively high persistence constants $\phi \approx 0.98$. When $\phi = 0.98$, pooling to depth $d = 100$ gives rise to residuals that account for $\phi^d \approx 0.13$ of the weight of the metric; and hence, if the metric score summed over the within-pool documents is (say) 0.3, then the unjudged documents might lift the score to 0.43. In practice such large jumps are uncommon, but as we have shown in Section 3, they can definitely occur as shallow-pool judgments are extended to deeper-pool evaluations.

In the absence of the bounds provided by residuals, we have sought to anticipate the behavior of recall-based metrics as judgment pools are deepened, and demonstrated that AP and NDCG scores tend to decrease as uncertainty is removed. We also sought a connection between the extent of any particular AP or NDCG score change as the judgments were deepened, and the size of the approximating RBP residual associated with the shallower evaluation, but found only weak correlations. This might be caused by factors that we have not controlled for in our analysis, or it might be that residuals – like statistical p-values, as we have also considered in our experiments – have little connection with metric consistency.

Overall, our experiments have demonstrated an important risk of the current IR evaluation process that has implications for reproducibility: statistical significance tests do not reflect the degree of uncertainty in per-topic point estimates that can arise from the presence of unjudged documents in a ranked results list. Moreover, based on the relationships we have documented between high-ϕ RBP scores and recall-based metrics such as AP and NDCG, those uncertainties can be non-trivial. When weighted-precision metrics such as RBP are used, we thus argue that residuals should always be presented in addition to statistical test values, as a secondary indicator of score consistency. Where recall-based metrics are the preferred choice, score behavior is less predictable, and no clear relationship with residuals was established. Nevertheless, as an adjunct to statistical significance tests, we still recommend that researchers provide information in regard to a high-ϕ RBP residuals, to help the reader assess the reliability of their results. Where such residuals cannot be computed for some reason, we recommend as an absolute minimum that authors be encouraged to report the fraction of unjudged documents among the top k documents for each topic, for some appropriate value of k (perhaps the limit on the evaluation depth, such as when NDCG@k is being computed), as a routine part of their experimental results presentation.

These steps will enhance awareness of the issues caused by finite-judgment processes; will promote clearer understanding of the measurement uncertainties that may be present in effectiveness-based experimental evaluation results; and will, ultimately, lead to experimental outcomes that are more reliable and hence more reproducible.

ACKNOWLEDGMENTS

The authors thank the referees for their support and their helpful feedback and suggestions. This work was supported by the Australian Research Council (DP180102687) and by a Google Faculty Research Grant.

REFERENCES

[1] J. A. Aslam, V. Pavlu, and R. Savell. 2003. A unified model for metasearch, pooling, and system evaluation. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 484–491.

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

[3] P. Bailey, A. Moffat, F. Scholer, and P. Thomas. 2016. UQV100: A test collection with query variability. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 725–728. Public data: http://dx.doi.org/10.4225/49/5726E597B8376.

[4] C. Buckley, D. Dimmick, I. Soboroff, and E. M. Voorhees. 2007. Bias and the limits of pooling for large collections. *Information Retrieval* 10, 6 (2007), 491–508.

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

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

[7] B. Carterette. 2007. Robust test collections for retrieval evaluation. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 55–62.

[8] B. Carterette, J. Allan, and R. K. Sitaraman. 2006. Minimal test collections for retrieval evaluation. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 268–275.

[9] O. Chapelle, D. Metzler, Y. Zhang, and P. Grinspan. 2009. Expected reciprocal rank for graded relevance. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 621–630.

[10] N. Ferro. 2017. What does affect the correlation among evaluation measures? *ACM Trans. on Information Systems* (2017). In press, online version accessed 28 Aug 2017.

[11] N. Fuhr. 2017. Some common mistakes in IR evaluation, and how they can be avoided. *SIGIR Forum* 51, 3 (2017), 32–41.

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

[13] A. Lipani, M. Lupu, J. Palotti, G. Zuccon, and A. Hanbury. 2017. Fixed budget pooling strategies based on fusion methods. In *Proc. Symp. Applied Computing (SAC)*. 919–924.

[14] A. Lipani, J. R. M. Palotti, M. Lupu, F. Piroi, G. Zuccon, and A. Hanbury. 2017. Fixed-cost pooling strategies based on IR evaluation measures. In *Proc. European Conf. in Information Retrieval (ECIR)*. 357–368.

[15] D. E. Losada, J. Parapar, and A. Barreiro. 2017. Multi-armed bandits for adjudicating documents in pooling-based evaluation of information retrieval systems. *Information Processing & Management* 53, 5 (2017), 1005–1025.

[16] X. Lu, A. Moffat, and J. S. Culpepper. 2016. The effect of pooling and evaluation depth on IR metrics. *Information Retrieval* 19, 4 (2016), 416–445.

[17] X. Lu, A. Moffat, and J. S. Culpepper. 2017. Can deep effectiveness metrics be evaluated using shallow judgment pools?. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 35–44.

[18] S. Mizzaro. 1997. Relevance: The whole history. *Journal of the American Society for Information Science and Technology* 48, 9 (1997), 810–832.

[19] A. Moffat. 2013. Seven numeric properties of effectiveness metrics. In *Proc. Asia Information Retrieval Societies Conf. (AIRS)*. 1–12.

[20] A. Moffat, P. Bailey, F. Scholer, and P. Thomas. 2017. Incorporating user expectations and behavior into the measurement of search effectiveness. *ACM Trans. on Information Systems* 35, 3 (2017), 24:1–24:38.

[21] A. Moffat, P. Thomas, and F. Scholer. 2013. Users versus models: What observation tells us about effectiveness metrics. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 659–668.

[22] A. Moffat, W. Webber, and J. Zobel. 2007. Strategic system comparisons via targeted relevance judgments. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 375–382.
[23] A. Moffat and J. Zobel. 2008. Rank-biased precision for measurement of retrieval effectiveness. *ACM Trans. on Information Systems* 27, 1 (2008), 2.1–2.27.
[24] S. D. Ravana and A. Moffat. 2009. Score aggregation techniques in retrieval experimentation. In *Proc. Australasian Database Conf. (ADC)*. 59–67. http://crpit.com/confpapers/CRPITV92Ravana.pdf
[25] S. Robertson. 2006. On GMAP: And other transformations. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 78–83.
[26] T. Sakai. 2004. New performance metrics based on multigrade relevance: Their application to question answering. In *Proc. NII Testbeds and Community for Information Access Research (NTCIR)*.
[27] T. Sakai. 2006. Evaluating evaluation metrics based on the bootstrap. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 525–532.
[28] T. Sakai. 2007. Alternatives to BPref. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 71–78.
[29] T. Sakai. 2008. Comparing metrics across TREC and NTCIR: The robustness to pool depth bias. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 691–692.
[30] T. Sakai. 2008. Comparing metrics across TREC and NTCIR: The robustness to system bias. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 581–590.
[31] T. Sakai. 2016. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 5–14.
[32] T. Sakai and N. Kando. 2008. On information retrieval metrics designed for evaluation with incomplete relevance assessments. *Information Retrieval* 11, 5 (2008), 447–470.
[33] M. Sanderson. 2010. Test collection based evaluation of information retrieval systems. *Foundations & Trends in Information Retrieval* 4, 4 (2010), 247–375.
[34] D. Sheskin. 1997. *Handbook of Parametric and Nonparametric Statistical Procedures*. CRC Press.
[35] M. D. Smucker, J. Allan, and B. Carterette. 2007. A comparison of statistical significance tests for information retrieval evaluation. In *Proc. ACM International Conf. on Information and Knowledge Management (CIKM)*. 623–632.
[36] K. Spärck Jones and C. J. Van Rijsbergen. 1975. Report on the need for and provision of an 'ideal' information retrieval test collection. *Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5266* (1975).
[37] E. M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. *Information Processing & Management* 36, 5 (2000), 697–716.
[38] E. M. Voorhees. 2004. Overview of the TREC 2004 robust retrieval track. In *Proc. Text Retrieval Conf. (TREC)*. http://trec.nist.gov/pubs/trec13/papers/ROBUST.OVERVIEW.pdf NIST Special Publication 500-261.
[39] E. M. Voorhees. 2004. Overview of TREC 2004. In *Proc. Text Retrieval Conf. (TREC)*. http://trec.nist.gov/pubs/trec13/papers/OVERVIEW13.pdf NIST Special Publication 500-261.
[40] E. M. Voorhees and D. Harman. 1998. Overview of the seventh Text REtrieval Conference. In *Proc. Text Retrieval Conf. (TREC)*. http://trec.nist.gov/pubs/trec7/papers/overview_7.pdf.gz NIST Special Publication 500-242.
[41] E. M. Voorhees and D. Harman. 1999. Overview of the eighth Text REtrieval Conference. In *Proc. Text Retrieval Conf. (TREC)*. http://trec.nist.gov/pubs/trec8/papers/overview_8.pdf NIST Special Publication 500-246.
[42] E. M. Voorhees and D. K. Harman (Eds.). 2005. *TREC: Experiment and Evaluation in Information Retrieval*. MIT Press.
[43] W. Webber, A. Moffat, and J. Zobel. 2010. The effect of pooling and evaluation depth on metric stability. In *Proc. Wkshp. Evaluating Information Access (EVIA)*. 7–15. http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings8/EVIA/03-EVIA2010-WebberW.pdf
[44] W. Webber, A. Moffat, and J. Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. on Information Systems* 28, 4 (2010), 20.1–20.38.
[45] Z. Yang, A. Moffat, and A. Turpin. 2016. How precise does document scoring need to be?. In *Proc. Asia Information Retrieval Societies Conf. (AIRS)*. 279–291.
[46] E. Yilmaz, J. A. Aslam, and S. Robertson. 2008. A new rank correlation coefficient for information retrieval. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 587–594.
[47] J. Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In *Proc. ACM Conf. on Research and Development in Information Retrieval (SIGIR)*. 307–314.

Received October 2017; revised April 2018; accepted July 2018