In the context of depth-\( k \) pooling for constructing web search test collections, we compare two approaches to ordering pooled documents for relevance assessors: The prioritisation strategy (PRI) used widely at NTCIR, and the simple randomisation strategy (RND). In order to address research questions regarding PRI and RND, we have constructed and released the WWW3E8 dataset, which contains eight independent relevance labels for 32,375 topic-document pairs, i.e., a total of 259,000 labels. Four of the eight relevance labels were obtained from PRI-based pools; the other four were obtained from RND-based pools. Using WWW3E8, we compare PRI and RND in terms of inter-assessor agreement, system ranking agreement, and robustness to new systems that did not contribute to the pools. We also utilise an assessor activity log we obtained as a byproduct of WWW3E8 to compare the two strategies in terms of assessment efficiency. Our main findings are: (a) The presentation order has no substantial impact on assessment efficiency; (b) While the presentation order substantially affects which documents are judged (highly) relevant, the difference between the inter-assessor agreement under the PRI condition and that under the RND condition is of no practical significance; (c) Different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition; and (d) PRI-based relevance assessment files (qrels) are substantially and statistically significantly more robust to new systems than RND-based ones. Finding (d) suggests that PRI helps the assessors identify relevant documents that affect the evaluation of many existing systems, including those that did not contribute to the pools. Hence, if researchers need to evaluate their current IR systems using legacy IR test collections, we recommend the use of those constructed using the PRI approach unless they have a good reason to believe that their systems retrieve relevant documents that are vastly different from the pooled documents. While this robustness of PRI may also mean that the PRI-based pools are biased against future systems that retrieve highly novel relevant documents, one should note that there is no evidence that RND is any better in this respect.

CCS Concepts: • Information systems → Test collections; Relevance assessment; Retrieval effectiveness;

Additional Key Words and Phrases: Information retrieval, pooling, relevance assessments, test collections, web search

ACM Reference format:
Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents?. ACM Trans. Inf. Syst. 40, 4, Article 76 (January 2022), 35 pages.
https://doi.org/10.1145/3494833
1 INTRODUCTION

Decades after the Cranfield II experiments based on relevance assessments in the 1960s [7, 8] and the proposal of pooling-based “ideal” test collections in the 1970s [42, 43], offline information retrieval system evaluation using pooling-based test collections still remains vital for providing researchers with insight into why some retrieval methods work, while others do not, and for helping them advance the state-of-the-art by building on top of that knowledge. More specifically, depth-\(k\) pooling, the approach of taking top \(k\) documents from each participating run to form a pool for relevance assessments [15, 30], has played a central role in collecting relevance assessments for test collections with large document collections, ever since the advent of TREC\(^1\) in the early 1990s. The present study concerns how the pooled documents are ordered and presented to the relevance assessors; in the context of depth-\(k\) pooling, we address a few research questions that have been debated over the past two decades or so, as discussed more specifically below.

While many IR tasks still rely on depth-\(k\) pooling, the pooled documents are ordered for the assessors differently at different tasks. In an article given at CLEF\(^2\) 2001, Voorhees explains the early TREC approach [46, p. 357]: “Each pool is sorted by document identifier so assessors cannot tell if a document was highly ranked by some system or how many systems (or which systems) retrieved the document.” An almost identical comment by Harman can be found in the TREC book published in 2005 [15, p. 33]. The philosophy there is to randomise the document judging order to avoid judgement biases: for example, if the assessor knows that a document was retrieved by many participating systems at top ranks, then this knowledge might make them overestimate its relevance. We call this the randomisation approach (or RND for brevity). Note that we do not take of the above advice of “sorting by document number” literally: if the publication date is embedded in the document identifier, then sorting by document ID would mean sorting by time, which is not what we want. Similarly, as Damessie et al. [10] have observed, if the target document collection consists of multiple subcollections and the document IDs contain different prefixes accordingly, such a sort would actually cluster documents by source, which again is not what we want.

Interestingly, many pooling-based IR tasks of NTCIR\(^3\) do not follow the above advice from TREC. In fact, they do the exact opposite. At CLEF 2003, Kando explained [17, p. 38]: “Pooled document lists to be judged are sorted in descending order of likelihood of being relevant (not the order of the document IDs)” and remarks that the NTCIR’s decision was “based on comparative tests and interviews with assessors.” However, the details were never published. In 2008, Sakai et al. [33] introduced a specific implementation of NTCIR’s ordering approach, now widely used at NTCIR with the NTCIRPOOL tool [30]\(^4\): the pooled documents are sorted by pseudo relevance, where the first sort key is the number of runs containing the document at or above the pool depth \(k\) (the larger the better), and the second sort key is the sum of ranks of that document within those runs (the smaller the better). Thus, documents that were retrieved by many systems at high ranks are prioritised. This NTCIR approach, hereafter referred to as PRI, aims at letting the assessors go through the documents roughly in decreasing order of relevance so that they can quickly form an idea as to what constitutes a relevant document, and thereby enhance assessment efficiency and inter-assessor agreement [38]. To date, however, there is no concrete evidence that supports the above claims.

RND and PRI are probably the two most simple and widely-used document ordering strategies for constructing depth-\(k\) pool-based IR test collections. Even after the past two decades or so\(^5\); however, the IR research community has yet to reach a consensus as to what advantages each

\(^1\)Text Retrieval Conference [15].
\(^2\)Cross-Language Evaluation Forum, now known as Conference and Labs of the Evaluation Forum [13].
\(^3\)National Institute of Informatics (NII) Testbeds and Community for Information access Research [35].
\(^4\)http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html.
\(^5\)The first NTCIR conference (workshop) was held in 1999 [35].
of these two strategies actually offer. Hence, the present study addresses the following research questions for PRI and RND.

**RQ1:** Which strategy enables more efficient relevance assessments?

**RQ2:** Which strategy enables higher inter-assessor agreements?

**RQ3:** Which strategy enables more stable system rankings across different versions of qrels files?

**RQ4:** Which strategy is more robust to the evaluation of systems that did not contribute to the pools?

In order to address research questions regarding PRI and RND, we have constructed and released the WWW3E8 dataset [36], which contains eight independent relevance labels for 32,375 topic-document pairs, i.e., a total of 259,000 labels. Four of the eight relevance labels were obtained from PRI-based pools; the other four were obtained from RND-based pools. Using WWW3E8, we compare PRI and RND in terms of inter-assessor agreement, system ranking agreement, and robustness to new systems that did not contribute to the pools. We also utilise an assessor activity log we obtained as a byproduct of WWW3E8 to compare the two strategies in terms of assessment efficiency. Our main findings are: (a) The presentation order has no substantial impact on assessment efficiency; (b) While the presentation order substantially affects which documents are judged (highly) relevant, the difference between the inter-assessor agreement under the PRI condition and that under the RND condition is of no practical significance; (c) Different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition; and (d) PRI-based relevance assessment files (qrels) are substantially and statistically significantly more robust to new systems than RND-based ones. Finding (d) suggests that PRI helps the assessors identify relevant documents that affect the evaluation of many systems, including those that did not contribute to the pools. Hence, if researchers need to evaluate their current IR systems using legacy IR test collections, we recommend the use of those constructed using the PRI approach unless they have a good reason to believe that their systems retrieve relevant documents that are vastly different from the pooled documents. Put another way, if the researchers believe that their new systems return search results that are reasonably similar to existing systems, then a PRI-based test collection is recommended; otherwise the new systems may be heavily underrated. While this robustness of PRI may also mean that the PRI-based pools are biased against future systems that retrieve highly novel relevant documents, one should note that there is no evidence that RND is any better in this respect.

The present study extends our SIGIR 2021 resource article that introduced the WWW3E8 dataset [36], by reporting on extensive experiments conducted by utilising WWW3E8. We acknowledge that Sections 2 and 3 basically duplicate the content of our resource article: This is to ensure that the present study is self-contained as a journal article. The major component of the present article is the examination of the four research questions mentioned above. We also acknowledge that the relevance assessments of the WWW3E8 dataset were obtained from bronze assessors [3], i.e., non-experts who are not the creators of the search topics (See Section 2); we shall discuss how we plan to follow up on this limitation in Section 9.

The present study is also a follow-up study of the work by Sakai and Xiao [38], who tried to address **RQ1** and **RQ2** but reported that the statistical powers of their experiments were too low. The design of our WWW3E8 dataset was directly based on the sample size considerations discussed by Sakai and Xiao. We shall discuss the differences between their work and ours in Section 2.3.

---

6 Available at https://waseda.box.com/WWW3E8dataset.
2 PRIOR ART

Test collection-based evaluations of IR systems depend on human relevance assessments and therefore ensuring the reliability of the assessments as the ground truth is of utmost importance to the IR community. Accordingly, there is a large body of work on the reliability of relevance assessments. For example, Voorhees [45] demonstrated that, while different assessors rate the same documents somewhat differently, when we rank systems by mean effectiveness scores, the system rankings are quite robust to the change in the set of relevance assessments. Bailey et al. [3] examined the effect of assessor expertise on ranking systems, by considering three types of relevance assessors: gold (topic originators), silver (task experts who are not topic originators), and bronze (who are neither). From this viewpoint, Cranfield II [7, 8], the first IR experiment that involved relevance assessments, involved gold assessors (in addition to students); the present study concerns bronze assessors, as we shall discuss in Section 3.3. Below, we focus our attention on prior art on document selection strategies for pooling (Section 2.1), and those on document ordering for relevance assessors (Section 2.2). Regarding document ordering, Section 2.3 specifically discusses the work of Sakai and Xiao [38], as their sample size considerations were utilised for the design of our WWW3E8 dataset; we also clarify the differences between their work and ours.

2.1 Alternatives to Depth-\(k\) Pooling

Although several alternatives to depth-\(k\) pooling have been proposed, it is probably fair to say that none is as widely-used. In 1998, Zobel [47] proposed to allocate more judging resources to more promising topics, while Cormack et al. [9] proposed to allocate more judging resources to more promising runs. TREC did not adopt these methods for fear of introducing bias [46]. The TREC Million Query Track [1, 4] used the Minimal Test Collections (MTC) and statAP methods. MTC iteratively orders documents to be judged according to how much information they provide about a difference in average precision; statAP samples documents to be judged according to a sampling distribution that tries to place higher weights on relevant documents. More recently, the TREC 2017 Common Core Track [2] re-examined the problem of how best to go beyond depth-\(k\) pooling. They adopted a version of the MaxMean method of Losada, Parapar, and Barreiro [20], which dynamically selects which run to process based on the judgements so far; this is similar in spirit to the aforementioned method of Cormack et al. in that it tries to focus judging resources on those runs that continue to contribute relevant documents. The Common Core Track 2017 overview article mentions judgement bias, i.e., the bias caused by presenting top-ranked documents to the assessors first, and run bias, i.e., underestimating runs that do not contribute many relevant documents early in their rankings. While the overview article reports that there was no indication of run bias in their experiments, judgement bias was not tested.

We are primarily interested in depth-\(k\) pooling because our view is that, generally speaking, dynamic document selection approaches have a few practical inconveniences when compared to depth-\(k\) pooling. Firstly, they require that once a relevance assessment is made, that is final. In contrast, with depth-\(k\) pooling, assessors can correct their judgements at any time in any order, as long as the judgement interface allows; the present study actually examines how often assessors correct their labels. Secondly, dynamic approaches complicate logistics: For example, compared to depth-\(k\) pooling, it may be more difficult to anticipate the workload per assessor, and to let multiple assessors handle the same topic. In short, we view the static nature of depth-\(k\) pooling as a strength.

In the aforementioned Common Core Track [2], the track coordinators point out that it is important to give each assessor an initial burn-in period, in which the assessor learns about the topic and optionally makes changes to their own initial relevance assessments; hence, the coordinators
implemented a hybrid approach where the assessor initially processes a traditional depth-10 pool and then moves to the Max-Mean-based dynamic judging phase. That is, even this TREC track actually relies on depth-k pooling. We also note that the aim of the burn-in period resembles NTCIR’s motivation for prioritising the pooled documents (See Section 1).

The recent work of Lipani et al. [19] provides a comprehensive study on alternatives to depth-k pooling, including adaptive (i.e., dynamic) methods such as the ones discussed above [20] as well as nonadaptive methods adopted from ranking fusion, i.e., computing pseudo relevance scores based on multiple ranked lists [34]. For nonadaptive settings, Lipani et al. recommend using the maximum score obtained by a document across all runs as an alternative to depth-k pooling. However, this is beyond the scope of the present study.

2.2 Document Ordering for Assessors

Document judging order was a concern even before TREC. For example, in 1988, Eisenberg and Barry [12] reported on a small-scale experiment where 15 document descriptions for a single topic were presented (on article) to each assessor in either increasing or decreasing order of relevance, and the assessors were asked to rate the documents on a 7-point scale. They observed a hedging phenomenon: The assessors were reluctant to label the early documents with very high or very low scores, because they might want to reserve these extreme scores for later documents. Thus, for example, when the documents were presented in decreasing order of relevance, the assessors tended to underestimate the relevance of the early documents. Based on their results, they caution against judgement bias and recommend randomising the presentation order. In 2004, Huang and Wang [16] reported on a similar experiment with similar recommendations, but varied the number of documents to judge: \( d = 5, 15, 30, 45, 60, 75 \). They reported that order effects were not observed when \( d = 5 \) and \( d = 75 \), and conjectured that the latter result may be due to fatigue: “an excess of documents simply exhausts the subjects with the toil and leads to responding without careful considerations.”

We do not consider the above studies as directly comparable to our work for the following reasons.

— They used a 7-point scale ratings, which is probably the primary reason for the hedging phenomenon: It is probably difficult to give a 7-point label or a 1-point label to an early document without having seen the other documents. In contrast, our relevance assessment task is simpler, as the choices are highly relevant, relevant, or nonrelevant (and error) as described in Section 3.2.

— Their relevance assessors examined the documents provided in article form (even in the 2004 article of Huang and Wang [16]); we use a web browser-based interface, which allows the assessors to go back and forth on the document list and even correct their relevance labels if they wanted to.

— Their experimental designs are based on the implicit assumption that the RND-based relevance assessments are perfect and correct: That is, they obtained the relevance assessments first based on the RND document ordering, and then sorted the judged documents based on the RND-based relevance levels to examine the order effect. In contrast, we make no such assumption: we directly compare the RND and PRI conditions and study which documents are judged relevant, as well as the outcomes of these differences.

— Their experiments relied on a single search topic; we have 160 different topics.

---

7 Eisenberg and Barry [12] considered magnitude estimation (See also Turpin et al. [44]) as an alternative to 7-point ratings; this is beyond the scope of our study.
Note also that while Huang and Wang [16] discussed the possibility of fatigue with \( d = 75 \) documents (in article form), modern relevance assessments (with documents presented on a computer screen) typically involves hundreds of documents per topic. In our study, each assessor handled 53–54 topics, where each topic has about 202 documents on average. While fatigue may well be playing a part in our experiments, this is beyond the scope of our study, as how each assessor works (e.g., how many topics or documents they process each day) during the two months that they were given completely at their discretion (See Section 3.2).

In 2013, Scholer et al. [41] studied the effect of the overall relevance of early documents on the 4-point assessor ratings of later documents, using three topics from TREC and 48 documents per topic. The first 20 documents presented to the assessors were called the Prologue; the other 28 were called the Epilogue. They controlled the quality of the Prologue based on existing relevance assessments (treated as the gold standard).\(^8\) By comparing across the three Prologues conditions (high/medium/low relevance), they observed an effect similar to the hedging phenomenon of Eisenberg and Barry [12], and argued that “people’s internal relevance models are impacted by the relevance of the documents they initially view and that they can re-calibrate these models as they encounter documents with more diverse relevance scores.”

In 2018, as a follow up to an earlier study from 2016 [11], Damessie et al. [10] compared three document ordering methods: the aforementioned prioritisation by NTCIRPOOL, randomised ordering, and their own method, which presents document blocks, where each block contains likely nonrelevant documents followed by a single pseudo relevant document. While this third method is beyond the scope of the present study as it has not yet been tested in actual evaluation venues, their results suggested that prioritisation achieves a higher inter-assessor agreement than randomisation. However, it should be noted that their experiments relied on only 240 topic-document pairs: eight topics (4 from TREC-7, 4 from TREC-8), each with 30 pooled documents.\(^9\) In contrast, the present study utilises our WWW3E8 dataset with 32,375 topic-document pairs and \( 8 \times 32,375 = 259,000 \) relevance labels, and therefore is far larger in scale. Note also that assessor efficiency was outside the scope of the study by Damessie et al. [10].

As we have mentioned in Section 1, Damessie et al. [10] pointed out that sorting pooled documents by document IDs can cluster documents by source because the source IDs are often encoded as document ID prefixes, and that this can inadvertently produce small clusters of relevant documents for the assessors. For this reason, they replaced their earlier approach of sorting by document IDs [11] with randomisation. The present study follows suit and compare the RND and PRI approaches.

Also in 2018, Losada, Parapar, and Barreiro [21] reported on a study on when to stop judging documents to reduce the assessment cost, under the premise that pooled documents are ranked by a kind of pseudo relevance. They remark: “Although there is still room for debate, we believe that a relevance-based ordering of assessments should not be an obstacle in practice.” Their view appears to be generally in line with the PRI approach.

### 2.3 Sakai and Xiao: Our Pilot Study

This section discusses the work of Sakai and Xiao [38] who also tried to address RQ1 (efficiency) and RQ2 (inter-assessor agreement) and clarifies our new contributions.

---

\(^8\) Unlike the earlier studies of Eisenberg and Barry [12] and Huang and Wang [16], the work of Scholer et al. used a relevance assessment interface as in our study.

\(^9\) Note that “30 pooled documents” is not the same as depth-30. In general, a depth-30 pool from multiple runs would give us many more than 30 documents. Our depth-15 pools gave us over 200 documents per topic (See Section 3.2).
The main differences between Sakai and Xiao’s work and ours are as follows:

— The statistical powers of their experiments were very low: below 18% in their efficiency experiments (using data from the NTCIR-14 WWW-2 task [22]) and about 32% in their inter-assessor agreement experiments (using data from the NTCIR-13 WWW-1 task [23]). Based on these results, they recommended creating 160 topics where each topic has 4 RND-based relevance assessments and 4 PRI-based relevance assessments. We constructed the WWW3E8 dataset by following exactly this recommendation to conduct experiments with higher statistical power. As a result, our efficiency experiment achieves over 65% statistical power for paired t-tests with four (out of five) of our efficiency criteria (as discussed in Section 4); our inter-assessor agreement also achieves up to about 91% statistical power for paired t-tests with mean weighted kappa [30] (as discussed in Section 5). In short, we have successfully utilised the work of Sakai and Xiao as a pilot study.

— Although they also addressed RQ1, their efficiency experiment had only one RND-based qrels and one PRI-based qrels per topic (with 80 topics and 19 runs). In contrast, by utilising four RND-based and four PRI-based qrels per topic (with 160 topic and 36 runs), we compared not only RND with PRI but also different versions of qrels within the RND and PRI conditions. Our work reports for the first time that the choice between RND and PRI probably has no substantial impact on assessor efficiency.

— Although they also addressed RQ2, their inter-assessor agreement experiment had only two RND-based qrels and two PRI-based qrels per topic (with 50 topics and 13 runs), and they examined the agreement between every pair of qrels in terms of weighted kappa. In contrast, using our four RND-based and four PRI-based qrels per topic (with 160 topic and 36 runs), we were able to focus on the general trends of agreement within each condition and across the two conditions in term of Krippendorff’s alpha. (We also discuss the largest kappa among all pairs of qrels to check that sufficient statistical power for kappa has been achieved.) Our work reports for the first time that the choice between RND and PRI probably has no substantial impact on assessor agreement either.

— Although they also reported on a preliminary system ranking agreement experiment (which overlaps with our RQ3), they warned thus: “the results reported in this section should be considered preliminary due to the small number of runs.” Again, they only had 13 runs evaluated on 50 topics. In contrast, because we have as many as $(4 \times 3/2) = 6$ RND-RND comparisons, six PRI-PRI comparisons, and $(4 \times 4) = 16$ RND-PRI comparisons, all based on ranking 36 runs, we are able to compare mean system ranking agreements in statistical terms. Our results show for the first time that RND-based and PRI-based system rankings are substantially different, and we clarify what this implies.

— They did not address our RQ4 (robustness to new systems), which is of utmost practical importance.

3 CONSTRUCTING THE WWW3E8 DATASET

3.1 NTCIR-15 WWW-3 English Subtask

This section briefly describes the English subtask of the NTCIR-15 WWW-3 task [37], as WWW3E8 was constructed while we served as a subgroup of the organisers for this task. WWW-3 was a traditional adhoc web search task, which offered Chinese and English subtasks. The test topics were released in March 25, 2020; the run submission deadline was May 31, 2020; the task was concluded at the NTCIR-15 conference in December 2020. The target corpus for the English subtask was cluweb12-B13.10 This subtask received 37 runs from 9 teams.

---

10https://lemurproject.org/cluweb12/.
Fig. 1. Topics 0001 (from the WWW-2 topic set) and 0101 (from the new WWW-3 topic set).

The WWW-3 English subtask participants were required to process 160 topics. The first 80 topics (0001–0080) were constructed as the test topics at the NTCIR-14 WWW-2 task [23], and the participants were given access to the WWW-2 version of the qrels file for these topics prior to run submission. Hence, it was possible for participants to tune their systems with these topics. The other 80 topics (0101–0180) were created at WWW-3 [37]. The common topic set size (80 for both WWW-2 and WWW-3) was determined at the WWW-2 task [23] based on topic set size design [28, 29]: It was estimated that 80 topics were more than sufficient for ensuring 80% statistical power for any $t$-test at the 5% significance level where the true difference between two systems is 0.10 or larger in terms of normalised Expected Reciprocal Rank (nERR). nERR was the least statistically stable measure used in the task. We decided to conduct relevance assessments for all 160 topics from scratch, based on the sample size considerations of Sakai and Xiao [38] (See Section 2.3). More details on the qrels construction step will be given in Section 3.2.

All 160 topics are publicly available. Out of them, 30 of the WWW-2 topics and 30 of the WWW-3 topics originate from query logs of Sogou, a major Chinese search engine; they were manually translated into English by the WWW organisers. The remaining 100 topics were sampled from the AOL query log. The WWW organisers ensured that the topic sets represent primarily torso queries [6], as head queries such as “facebook.com” are less interesting for modern web search research. Figure 1 shows Topics 0001 and 0101 as examples: Each query (i.e., topic) contains a content (or title in TREC parlance) field and a description field. Both queries originate from Sogou. The descriptions were composed of the WWW organisers to back-fit the intent and context behind the query.

3.2 Constructing the Qrels Files

Based on the recommendation from Sakai and Xiao [38] (See Section 2.3), we decided to construct four versions of qrels based on RND pools, and another four based on PRI pools, for all of the aforementioned 160 topics. The construction procedure is described below.

Based on our budget, we formed depth-15 pools from the 37 submitted runs. This gave us 32,375 topic-document pairs to judge. Hereafter, we shall refer to topic-document pairs as topicdocs for brevity. The average pool size is 32,375/160 = 202.3. Two versions of pool files were created for each topic: One in which the document order is randomised (RND), and one based on NTCIRPOOL

---

11https://waseda.box.com/www2www3topics-E.
12https://www.sogou.com/.

ACM Transactions on Information Systems, Vol. 40, No. 4, Article 76. Publication date: January 2022.
Constructing eight independent versions of qrels (four RND-based and four PRI-based) meant that a total of $32,375 \times 8 = 259,000$ labels were required.

As relevance assessors, we hired 24 international-course (English-based programme) computer science students from our university.\(^{13}\) Out of them 10 were undergraduates; the rest were master students. We divided them at random into three groups, each consisting of eight assessors, in order to obtain a $24 \times 32,375$ relevance assessment label matrix that is structured as shown in Figure 2. Group A was given 54 topics (11,154 topicdocs), Group B was given 53 topics (10,838 topicdocs), and Group C was given the remaining 53 topics (10,383 topicdocs). Within each group, the topics and the pools (either RND or PRI) were randomly assigned, while making sure that every assessor experienced both RND and PRI conditions in a reasonably balanced manner.

Based on past statistics [38], we told the assessors that they are expected to spend 20 seconds on each document on average and that they will be paid on that basis. This meant that the total work hours of each assessor in Group A–C were 62, 61, and 58, respectively. We also told them in advance that inter-assessor agreement will be checked upon job completion for quality control. The assessors were not aware that we were experimenting with two document ordering strategies; they were only told that their relevant assessments will be used for web search evaluation. The hourly pay was 1,200 JPY. Hence, the total cost for constructing the label matrix was $(62 + 61 + 58) \times 8 \times 1,200 = 1,737,600$ JPY. The assessors were given two months (from June 15 to August 15) to complete the job; during this period, we sent them a reminder with each assessor’s progress statistics, approximately once every two weeks. All the assessors managed to meet the deadline.

We instructed the assessors to go through a two-page manual first. The manual concisely describes the relevance assessment task, and how to use our browser-based relevance assessment interface called PLY [22, 30]. Each assessor was given a user account to log on to PLY, and was allowed to work from any place where a desktop PC with internet access is available, at any time during the above two months.

Figure 3 shows a screenshot of the PLY interface where the RND-based pool file for Topic 0101 has been loaded. In principle, the assessor can judge documents in any order by clicking on a document in the left panel. In practice, however, they usually judge the documents from top to bottom, as judging one document automatically takes them to the next document. As can be seen, for each document, the assessors were required to choose from four labels by clicking on a button. These labels were defined in the aforementioned manual as follows:

**H.REL** highly relevant—It is **likely** that the user with the information need shown will find this page relevant.

**REL** relevant—It is **possible** that the user with the information need shown will find this page relevant.

---

\(^{13}\) Our labelling procedure conforms to the guidelines provided by the Office of Research Ethics of our university.
The fourth option is needed because the cluweb corpus has character encoding problems with some documents. We thus obtained 28,144 H.REL, 61,512 REL, 163,090 NONREL, and 6,254 ERROR labels in total. We then treated each H.REL label as 2-relevant (highly relevant), each REL label as 1-relevant (relevant), and each NONREL or ERROR label as a 0-relevant, to form 3-point graded relevance data.

Figure 4 shows the distribution of relevance labels thus obtained for each qrels file. It can be observed that the distributions are quite similar and there are no noticeable differences between PRI and RND.

### 3.3 Data Reliability

Our relevance assessors are students: They are neither topic originators nor topic experts (i.e., bronze assessors [3]). Chouldechova and Mease [5] report that query owners (i.e., topic originators) provide more valuable relevance assessments than non-owners for web search evaluation. Hence, while WWW3E8 is a large-scale relevance assessment data, the reliability of the labels may not be as high as those that gold assessors (i.e., topic originators) might have provided. To shed some light on this issue, this section discusses data reliability in terms of Krippendorff’s $\alpha$ [18, 30].

As illustrated in Figure 2, the entire label dataset is a $24 \times 32,375$ matrix containing 2’s (highly relevant), 1’s (relevant), 0’s, and NA’s (for Not Available). Each topicdoc has eight labels plus 16 NA’s, since three topic subsets were handled by Groups A, B, and C, respectively. First, from this original label matrix, we computed Krippendorff’s $\alpha$ for ordinal labels [18, 30], and obtained 0.288.
This seems reasonable since, just like Cohen’s $\kappa$ (which can only be used for two assessors), Krippendorff’s $\alpha$ represents agreement beyond chance.

Next, to examine whether each assessor did a conscientious job, we computed leave-one-out $\alpha$ scores \[40\]. For example, we replace all labels contributed by Assessor 01 in the $24 \times 32,375$ label matrix with NA’s, and then recompute the $\alpha$. If the $\alpha$ goes up as a result, that means that Assessor 1 was hurting the overall reliability of the original label matrix. Table 1 shows the leave-one-out $\alpha$ scores. For example, the result of replacing Assessor 01’s labels with NA’s is shown as “w/o A01.” It can be observed that there are no outliers. The results suggest that the reliability of the assessments from the 24 assessors are comparable to one another. We acknowledge, however, that the results do not guarantee the accuracy of the relevance assessments. We only claim that WWW3E8 should serve as a useful common ground for researchers to compare the RND and PRI strategies within our bronze assessor setting.

### 3.4 Runs and Score Matrices

WWW3E8 also contains the original 37 runs that contributed to the depth-15 pools and the topic-by-run score matrices for the measures used at the NTCIR-15 WWW-3 task, created with each of the eight qrels files. Along with the run files, WWW8E3 also contains a file called Eruns-1, which is a list of 36 run file names. One run has been excluded because this run does not represent a single system: The first 80 topics and the other 80 topics were processed by different systems \[37\]. Hence, for evaluation experiments, using this run is not recommended.

As with the official results of the NTCIR-15 WWW-3 task, we used the NTCIREVAL toolkit \[30\] to compute all evaluation measures at the measurement depth of 10. The official measures are the

---

\[14\] Some researchers use Fleiss’ $\kappa$ for quantifying the agreement between more than two assessors. However, it should be noted that Fleiss’ $\kappa$ was designed for nominal labels \[14\].

\[30\] \url{http://research.nii.ac.jp/ntcir/tools/ntcireval-en.html} (version 200626).
Table 1. Krippendorff’s α Scores (for Ordinal Data) When Each Assessor’s Labels are Replaced with “NA” from the Original 24 × 32,375 Assessor-by-topicdoc Label Matrix for the 160 Topics

| w/o A01 | w/o A02 | w/o A03 | w/o A04 | w/o A05 | w/o A06 | w/o A07 | w/o A08 |
|---------|---------|---------|---------|---------|---------|---------|---------|
| 0.287   | 0.287   | 0.290   | 0.296   | 0.293   | 0.290   | 0.293   | 0.287   |
| w/o A09 | w/o A10 | w/o A11 | w/o A12 | w/o A13 | w/o A14 | w/o A15 | w/o A16 |
| 0.288   | 0.285   | 0.286   | 0.281   | 0.283   | 0.283   | 0.281   | 0.285   |
| w/o A17 | w/o A18 | w/o A19 | w/o A20 | w/o A21 | w/o A22 | w/o A23 | w/o A24 |
| 0.294   | 0.289   | 0.290   | 0.294   | 0.289   | 0.289   | 0.287   | 0.289   |

For example, “w/o A01” means “when all labels of Assessor 01” are replaced with an “NA.” The α with the original matrix is 0.288.

Microsoft version of normalised Discounted Cumulative Gain (nDCG) [26], the cutoff-based Q-measure [26], nERR (See Section 3.1), and intentwise Rank-Biased Utility (iRBU) [31, 40]. According to the experiments by Sakai and Zeng [39, 40], nDCG and iRBU outperformed other well-known measures in terms of agreement with users’ SERP preferences.

As each of the qrels files contains 3-point relevance levels (0, 1, and 2), we let the gain values be $2^1 - 1 = 1$ for 1-relevant documents, and $2^2 - 1 = 3$ for 2-relevant documents (i.e., exponential gain value setting). As for the patience parameter $p$ of iRBU, we used $p = 0.99$, the default value set in NTCIREVAL based on the results of Sakai and Zeng [39, 40].

4 RQ1: EFFICIENCY

While constructing WWW3E8, we obtained assessor activity logs from the PLY relevance assessment interface (See Figure 3). Following Sakai and Xiao [38], we collected the following efficiency statistics for each topic-assessor pair (i.e., for each topic-qrels pair) to address RQ1, our assessment efficiency question.

- **TJ1D** Time to judge the first document.
- **TF1RH** Time to find the first relevant or highly relevant document.
- **TF1H** Time to find the first highly relevant document.
- **ATBJ** Average time between judging two documents.
- **NREJ** Number of times the label of a judged document is corrected to another label.

For **TJ1D**, **TF1RH**, and **TF1H**, times longer than 3 minutes were considered outliers and were replaced with an “NA,” as we cannot tell from the log whether the assessors were actually reading a document or doing something else. Similarly, for computing **ATBJ**, times longer than 3 minutes were excluded when computing the average.\(^{16}\) Note that **ATBJ** is the most direct measure of assessor efficiency.

Table 2 shows, for each qrels file (PRI1 through RND4), our five efficiency criteria averaged across the topics. Note that the sample sizes are much smaller than 160 for **TJ1D**, **TF1RH**, and **TF1H** because we removed every topic that resulted in an “NA” for at least one version of the qrels. Scores averaged over all four PRI (RND) qrels versions are also shown. For each efficiency criterion, as we have eight mean scores to compare, we conducted a paired Tukey HSD test [29] at the 5% significance level. For **TJ1D**, **TF1RH**, and **TF1H**, none of the pairwise differences is statistically significant. It can also be observed that the effect sizes (i.e., standardised mean differences) [29] are also small. For example, for **TJ1D** in Table 2, the difference between the largest and smallest

\(^{16}\) We decided to adopt this 3-minute threshold from Sakai and Xiao [38], after observing that the ATBJ statistics in our experiments are similar to theirs (about 14 seconds as shown in Table 2).
Table 2. Efficiency Comparison for the Eight Qrels Files (PRI1 Through RND4) for the 160 Topics

| Criterion | n  | RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 | $V_{E2}$ |
|-----------|----|------|------|------|------|------|------|------|------|---------|
| TJ1D      | 44 | 34.3 | 41.2 | 39.7 | 38.8 | 34.3 | 35.7 | 37.1 | 28.1 | 1200.5  |
| TF1RH     | 72 | 25.1 | 28.9 | 28.0 | 27.2 | 31.3 | 31.5 | 29.9 | 29.0 | 920.3   |
| TF1H      | 59 | 13.3 | 23.4 | 24.2 | 18.6 | 26.6 | 22.4 | 23.9 | 27.4 | 704.9   |
| ATBJ      | 160| 13.1 | 15.8 | 14.5 | 14.8 | 14.5 | 14.3 | 14.7 | 14.7 | 56.53   |
| NREJ      | 160| 8.41 | 6.51 | 6.11 | 3.99 | 6.79 | 7.31 | 5.99 | 3.82 | 107.4   |

TJ1D, TF1RH, TF1H, and TF1H are in seconds; NREJ represents the number of label corrections. For each efficiency criterion, a paired Tukey HSD test at the 5% significance level was conducted to compare every pair of means. All statistically significant differences are indicated in the table with the $p$-value ($p$) and the effect size $ES$. $V_{E2}$ denotes the two-way ANOVA residual variance for computing the effect sizes [29]. In each row, the largest and smallest among the eight values are shown in bold and italics, respectively.

means is RND2 – PRI4 = 41.2 – 28.1 = 13.1; if we convert this to a standardised mean difference using the residual variance shown in the table, we obtain $13.1/\sqrt{1200.5} = 0.377$. That is, even the largest observed difference is less than half a standard deviation apart. Similarly, the largest effect sizes for TF1RH and TF1H are 0.211 (PRI2 – RND1) and 0.532 (PRI4 – RND1), respectively.

The lack of statistical significance for TJ1D, TF1RH, and TF1H is due to the high residual variances as shown in the $V_{E2}$ column of Table 2. As mentioned above, the effect sizes for these criteria are very small, and are probably not practically significant. However, we do note that the average TF1RH values of PRI are all larger than those of RND, and that the largest TF1H value of PRI (PRI4; 27.4 seconds) is twice as large as that of the smallest TF1H value of RND (RND1; 13.3 seconds). Hence, we closely examined the largest difference observed for TF1RH (RND1 vs. PRI2) and that observed for TF1H (RND1 vs. PRI4), as discussed below.

Figure 5 visualises the per-topic differences between RND1 and PRI2 in terms of TF1RH. There are 30 topics above the horizontal axis (PRI2 is more efficient), 41 topics below it (RND1 is more efficient), along with one topic right on the axis. Thus, even from this most extreme case (i.e., RND1 vs. PRI2), it is difficult to say which document ordering strategy is more efficient.

Figure 6 visualises the per-topic differences between RND1 and PRI4 in terms of TF1H. Regarding this particular comparison, while there is no statistical significance ($p = 0.077$, hence not shown in Table 2), there is a weak indication that RND1 is on the whole more efficient than PRI4: there are 20 topics above the horizontal axis (PRI4 is more efficient), 36 topics below it (RND1 is more efficient), along with 3 topics right on the axis. The average TF1H value of PRI4 suffers mainly due to the topics indicated with balloons in Figure 6: For example, for Topic 0003 (“women’s clothing winter”), it took the assessor who used an RND pool found their first highly relevant documents in 7 seconds, while the one who used a PRI pool found their first one after 168 seconds. However, after a close examination, we found that this unusual efficiency of RND1 is probably

\[\text{This is consistent with the results of Sakai and Xiao [38], where RND slightly outperformed PRI in terms of TF1RH on average (26.7 seconds vs. 32.5 seconds; } \Delta = 0.172), \text{ although this difference was not statistically significant either.}\]
largely due to assessor variance rather than the document ordering strategy. More specifically, we found that Assessor 01 from Group A and Assessor 17 from Group C (see Figure 2) contribute more assessments to RND1 than the other members of their groups do: Assessor 01 was responsible for as many as 35 topics (out of 160) of RND1, while other members of Group A handled only 24–34 topics for RND1. In particular, for TF1RH, Assessor 01 was responsible for 21 topics out of 72; for TF1H, Assessor 01 was responsible for 12 topics out of 59. Similarly, Assessor 17 was responsible for 35 topics (out of 160) of RND1, while other members of Group C handled only 23–33 topics for

\[\Delta \text{TF1RH per topic: RND1 minus PRI2 (72 topics).}\]

\[\Delta \text{TF1H per topic: RND1 minus PRI4 (59 topics).}\]

\[18\text{Recall that Group A assessors handled 54 topics. Therefore, Assessor 01 used PRI-based pools for } 54 - 35 = 19 \text{ topics.}\]
Table 3. Mean Per-assessor Efficiency Statistics

| Assessor | TJ1D  | TF1RH | TF1H  | ATBJ  | NREJ  |
|----------|-------|-------|-------|-------|-------|
|          | RND   | PRI   | RND   | PRI   | RND   | PRI   | RND   | PRI   | RND   | PRI   |
| A01      | 22.9  | 31.1  | 18.7  | 28.1  | 11.3  | 23.1  | 10.8  | 8.3   | 5.46  | 7.32  |
| A02      | 40.1  | 33.6  | 33.1  | 26.2  | 16.6  | 19.1  | 16.4  | 14.7  | 2.18  | 3.85  |
| A03      | 27.0  | 28.4  | 18.8  | 25.7  | 13.8  | 15.8  | 12.4  | 14.1  | 3.20  | 3.90  |
| A04      | 24.0  | 20.9  | 21.5  | 19.0  | 19.0  | 11.8  | 11.0  | 11.8  | 4.88  | 4.10  |
| A05      | 28.0  | 21.6  | 22.1  | 14.5  | 13.1  | 18.1  | 18.8  | 2.40  | 3.85  | 2.69  |
| A06      | 50.7  | 36.5  | 17.3  | 24.8  | 12.2  | 11.1  | 16.1  | 16.1  | 3.46  | 2.60  |
| A07      | 28.0  | 23.7  | 22.4  | 9.8   | 22.0  | 11.8  | 12.0  | 3.25  | 3.00  | 8.17  |
| A08      | 24.9  | 48.3  | 16.5  | 30.1  | 14.9  | 15.4  | 14.5  | 4.88  | 4.10  | 8.92  |
| A09      | 47.1  | 43.4  | 41.1  | 32.7  | 13.9  | 10.4  | 15.4  | 6.78  | 6.00  | 20.44 |
| A10      | 48.3  | 37.2  | 12.4  | 27.5  | 13.9  | 7.4   | 11.4  | 6.7   | 3.76  | 5.50  |
| A11      | 34.9  | 29.1  | 26.5  | 26.0  | 15.0  | 30.3  | 14.9  | 15.4  | 7.36  | 5.00  |
| A12      | 60.8  | 54.5  | 31.1  | 34.4  | 13.3  | 19.8  | 15.3  | 4.17  | 6.48  | 6.48  |
| A13      | 23.5  | 26.9  | 16.8  | 24.9  | 13.4  | 17.5  | 11.6  | 3.92  | 5.69  | 19.43 |
| A14      | 41.9  | 54.3  | 16.6  | 45.7  | 15.8  | 33.5  | 13.7  | 4.17  | 6.48  | 28.11 |
| A15      | 77.1  | 49.6  | 73.7  | 41.5  | 47.7  | 40.3  | 37.2  | 36.3  | 3.96  | 2.17  |
| A16      | 68.6  | 58.7  | 44.7  | 49.5  | 24.3  | 27.6  | 24.3  | 2.48  | 3.23  | 14.2  |
| A17      | 54.2  | 37.1  | 40.6  | 43.7  | 32.4  | 28.5  | 16.4  | 19.43 | 28.11 |
| A18      | 56.1  | 40.1  | 53.5  | 28.6  | 13.4  | 12.2  | 5.33  | 2.35  | 2.35  | 14.2  |
| A19      | 57.9  | 38.3  | 28.9  | 36.3  | 19.8  | 26.9  | 12.4  | 3.88  | 3.18  | 11.75 |
| A20      | 31.7  | 17.3  | 18.8  | 14.2  | 16.0  | 17.0  | 9.5   | 8.04  | 11.75 |
| A21      | 46.4  | 41.2  | 22.2  | 34.9  | 16.0  | 31.7  | 14.9  | 6.88  | 4.07  | 6.88  |
| A22      | 41.6  | 41.5  | 29.6  | 34.2  | 16.3  | 30.3  | 14.2  | 3.29  | 2.86  | 4.17  |
| A23      | 26.9  | 32.8  | 26.1  | 27.8  | 14.2  | 28.1  | 14.3  | 3.09  | 2.53  | 3.09  |
| A24      | 46.3  | 42.0  | 37.9  | 31.5  | 20.7  | 18.0  | 8.6   | 7.8   | 3.70  | 5.17  |

Largest and smallest values in each column are shown in **bold** and in *italics*, respectively.

RND1. In particular, for **TF1 RH**, Assessor 17 was responsible for 16 topics out of 72; for **TF1H**, Assessor 17 was responsible for 17 topics out of 59. The four outliers indicated in Figure 6 are also due to these two assessors: Topics 0003 and 0104 were handled by Assessor 01, and Topics 0156 and 0157 were handled by Assessor 17.

To examine the behaviours of Assessors 01 and 17 more closely, Table 3 shows the mean per-assessor efficiency statistics for all five efficiency criteria and for each document ordering strategy, for every assessor. It can be observed that, under the RND condition, Assessor 01 is on average the fastest among the 24 assessors in terms of **TJ1D** (22.9 seconds, which is an average over the aforementioned 35 topics); Assessor 01 is also the third fastest in terms of **TF1H** (11.3 seconds). As for Assessor 17, it can be observed that this assessor corrected the labels very often, and is not a particularly fast assessor. Nevertheless, since the fact remains that these two assessors are dominant factors of RND1, we cannot rule out the possibility that the differences in terms of **TF1 RH** and **TF1H** are primarily due to assessor variance rather than the choice of the document ordering strategy. Furthermore, even if the differences between RND and PRI in terms of these efficiency statistics are *real*, the reader should be reminded that the effect sizes are very small.

Table 2 also shows a few statistically significant differences for **ATBJ** and **NREJ**. Recall that **ATBJ** is the most direct criterion of assessor efficiency. Again, however, the effect sizes are very small. The statistically significant difference for **ATBJ** is between two versions of RND-based qrels.
Table 4. Achieved Statistical Powers if Paired \( t \)-tests are Used for the Largest Differences Shown in Table 2, with Sample Sizes Required for Achieving 70% Power

| Criterion | actual \( n \) | largest difference | \( t \) statistic | achieved power | \( n \) for 70\% power |
|-----------|----------------|-------------------|-----------------|----------------|---------------------|
| TJ1D      | 44             | RND2 > PRI4       | 2.42            | 65.6\%         | 49                  |
| TF1RH     | 72             | PRI2 > RND1       | 1.33            | 26.0\%         | 253                 |
| TF1H      | 59             | PRI4 > RND1       | 3.01            | 84.1\%         | 43                  |
| ATBJ      | 160            | RND2 > RND1       | 4.00            | 97.8\%         | 64                  |
| NREJ      | 160            | RND1 > PRI4       | 3.85            | 96.9\%         | 69                  |

Sakai’s R script `future.sample.pairedt` [27, 29] was used for the analysis.

(namely, RND1 and RND2) and is not interesting. As for NREJ, although the assessors involved in PRI4 corrected their labels statistically significantly less frequently compared to those involved in RND1, the effect size is only 0.443. Recall also the above discussion of assessor variance: Assessor 17, who is the second most frequent “label corrector” under the RND condition (See the NREJ (RND) column in Table 3), was responsible for as many as 35 topics and heavily affected the overall NREJ score of RND1 shown in Table 2 (8.41 times).

In summary, since none of the above differences in our efficiency criteria are of practical significance, our answer to RQ1 is: the choice of document ordering strategy (RND or PRI) has no substantial impact on assessor efficiency.

As was discussed in Section 2.3, we aimed for 70% statistical power (when paired \( t \)-tests are used) based on the sample size considerations of Sakai and Xiao [38]. Table 4 examines the actual achieved statistical powers for the largest differences observed in Table 2, using the R script `future.sample.pairedt` from Sakai [27, 29].19 For example, if we conduct a paired \( t \)-test with RND2 and PRI4 for the TJ1D row of Table 2,\(^{20}\) the achieved power is about 65.6\%, which is close to what we originally aimed for. It can be observed that, with the exception of TF1RH, we have managed to achieve high statistical powers. Finally, while our experiment is underpowered for TF1RH, whether this deserves further investigation with a larger sample is questionable, given the above discussions of small effect sizes and assessor variance.

5 RQ2: INTER-ASSESSOR AGREEMENT

We now utilise the WWW3E8 dataset to address RQ2 (Which document ordering strategy enables higher inter-assessor agreements?). We quantify the inter-assessor agreements under RND and PRI conditions using Krippendorff’s \( \alpha \) for ordinal classes [18, 30]. For example, to quantify the inter-assessor agreement under the RND condition, all labels in the original WWW3E8 matrix that were obtained under the PRI condition can be replaced with NA’s and then the \( \alpha \) can be recomputed, so that each topic/doc has only four labels instead of eight. Table 5 shows the results. It can be observed that while the \( \alpha \) scores for RND and PRI are very similar, they are much higher than the \( \alpha \) score for the original matrix (0.288). That is, while the labels within each document ordering strategy are similar to each other, the labels across the two strategies differ substantially. It is clear that the document ordering strategy substantially affects which documents are judged (highly) relevant. We shall provide an explanation for this in Section 6.2.

19 Available from https://waseda.box.com/SIGIR2016PACK.
20 Note that we do not actually discuss this \( t \)-test: instead, Table 2 uses the Tukey HSD test (i.e., a multiple comparison procedure) that utilises all eight versions of qrels.
Table 5. Overall Inter-assessor Agreement in Terms of Krippendorff’s $\alpha$ (for Ordinal Data) Based on the $24 \times 32,375$ Label Matrices for the 160 Topics

|        | All | RND | PRI |
|--------|-----|-----|-----|
| $\alpha$ | 0.288 | 0.433 | 0.423 |

The original matrix contains eight labels per topicdoc (four based on RND, four based on PRI); the RND and PRI matrices each contain four labels per topicdoc. All other cells are stuffed with “NA.”

Table 6. Mean Per-topic Krippendorff’s $\alpha$ (for Ordinal Data) Averaged Over the 160 Topics

|        | RND | PRI | $t$ statistic | $p$-value | Glass’s $\Delta$ | Achieved power | $n$ for 70% power |
|--------|-----|-----|---------------|-----------|------------------|----------------|-------------------|
| $\alpha$ | 0.293 | 0.279 | 0.949 | 0.344 | 0.0859 | 15.7% | $n = 1,098$ |

Each per-topic matrix contains four labels (either based on RND or PRI) per document. All other cells are stuffed with “NA.” The mean $\alpha$ when all eight labels are included in the matrix is 0.125. A paired $t$-test at the 5% significance level was conducted. Glass’s $\Delta$ [29] is based on the standard deviation from the RND data.

Table 7. Verifying the Achieved Statistical Power in Terms of Mean Kappa

|        | RND1 vs. RND2 vs. RND3 | $t$ statistic | $p$-value | Achieved power | $n$ for 70% power |
|--------|------------------------|---------------|-----------|----------------|-------------------|
| $\alpha$ | 0.175 | 0.141 | 3.33 | 0.00108 | 91.1% | 91 |

The above analysis computed a single $\alpha$ score for the entire matrix. In contrast, Table 6 compares the inter-assessor agreement under the RND and PRI conditions based on mean per-topic $\alpha$ scores, averaged over the 160 topics. According to a paired $t$-test, the difference between the RND and PRI conditions is not statistically significant. More importantly, the effect size (Glass’s $\Delta$, a form of standardised mean difference [29]) in terms of $\alpha$ is very small ($\Delta = 0.0859$), and power analysis asks for over 1,000 topics to achieve 70% statistical power for such a small effect size.$^{21}$ From these results, we conclude that even though the document ordering strategy substantially affects which documents are judged (highly) relevant, the difference between the inter-assessor agreement under the RND condition and that under the PRI condition is of no practical significance.

In Section 2.3, we discussed the sample size considerations of Sakai and Xiao [38], who recommended having 160 topics for achieving over 70% statistical power for mean kappas. Table 7 shows the achieved statistical power for the largest difference among mean weighted kappas of all pairs of qrels, namely, the difference between the kappa for RND1 and RND4, and that for RND2 and RND3. It can be observed that the achieved power is over 91%.

We also observe from Table 7 that the pairwise agreements in terms of kappa are lower compared to Sakai and Xiao [38, Table 3]: their kappa values lay between 0.247 (PRI1 vs. PRI2) and 0.341 (RND2 vs. PRI2). This suggests that there is considerable noise in the relevance assessments of the WWW3E8 dataset. Muraoka, Zeng, and Sakai [24] manually examined a small number of documents as a participating team of the NTCIR-15 WWW-3 task and also reported on some possible noise in the relevance assessments for the WWW-3 topics. Evaluating the accuracy of relevance assessments require gold relevance assessments [3], and is therefore left for future work (See Section 9).

$^{21}$ Sakai’s tool future.sample.pairedt [27] was used for the power analysis.
RQ3: SYSTEM RANKING AGREEMENT

6.1 System Ranking τ Results

Using the eight qrels files (PRI1 through RND4) available in WWW3E8, we now address **RQ3** (Which strategy enables more stable system rankings across different versions of qrels files?). More specifically, using each qrels file, we rank the 36 runs submitted to the NTCIR-15 WWW-3 task [37] with the official measures used in the task; recall that one run was excluded as described in Section 3.4. We then quantify the system ranking similarity with Kendall’s τ [26]. As we have mentioned in Section 3.4, the official measures used in the WWW-3 task are nDCG, Q-measure, nERR, and iRBU. These were computed using NTCIREVAL with an exponential gain value setting: 3 for highly relevant and 1 for relevant.

In this section as well as Section 7, evaluation measure scores are averaged over 147 topics rather than 160 topics for the following reasons. (1) There are a small number of cases where the qrels files (PRI1 through RND4) do not have any relevant documents. The union of such topics across all eight qrels files amounts to 8 topics (0012,0024,0026,0044,0063,0147,0169,0174). (2) In Section 7, where we discuss the subsets of the above qrels files to address **RQ4** (robustness to new systems), we lose 5 topics in addition for the same reason (0027,0060,0132,0153,0179). Hence, in order to evaluate the runs using a common topic set for every experimental condition, we average the evaluation measures over 160 – 8 – 5 = 147 topics. We have a total of 29,522 topicdocs from the depth-15 pools for these topics.

Table 8 shows the results of comparing all pairs of qrels versions in terms of τ. The trends are similar across all four evaluation measures, and are very clear. More specifically:

- The four PRI-based qrels files produce very similar system rankings (τ ≥ 0.841).
- The four RND-based qrels files produce moderately similar system rankings (0.495 ≤ τ ≤ 0.781).
- The RND-based rankings and the PRI-based ones are substantially different (τ ≤ 0.508).

The above three levels of system ranking agreement can be examined more closely as follows. From Table 8, we can compute, for each evaluation measure, a mean τ that represents the agreement within the RND condition by averaging the six values that compare two RND-based rankings. Similarly, we can obtain a mean τ within the PRI condition by averaging the six values that compare two PRI-based rankings. Finally, we can obtain a mean τ across the two conditions by averaging the 4 * 4 = 16 values that compare an RND-based ranking and a PRI-based ranking. To discuss the differences in means for these three cases, we can apply a Tukey HSD test for unpaired data at the 5% significance level [29].

Table 9 shows the results of the unpaired Tukey HSD test for each evaluation measure. Again, the results are similar for all four measures: the “Mean τ” columns show that system rankings within the PRI condition are very similar, those within the RND condition are less so, and that those across the two conditions are substantially different. As the “p-value” columns show, all of these differences in means are statistically highly significant. The “V<sub>E1</sub>” column shows the residual variance from one-way ANOVA for computing the effect sizes, since we are dealing with unpaired data here [29]. For example, the difference in mean nDCG between the within-RND condition (RND-RND) and the within-PRI condition (PRI-PRI) is 0.917 – 0.741 = 0.176; therefore, the effect size can be computed as 0.176/√0.00831 = 6.12. That is, the two means are about six standard deviations apart. We conclude that different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition. Moreover, as we have observed in Table 8, PRI-based rankings and RND-based rankings are substantially different from each other.
Table 8. System Ranking Agreement as Measured by Kendall’s $\tau$ between System Ranking Pairs ($n = 36$ Runs). According to the Four Official Measures of the NTCIR-15 WWW-3 Task, Averaged Over 147 Topics

(a) nDCG

| System   | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|----------|------|------|------|------|------|------|------|
| RND1     | 0.746| 0.752| 0.778| 0.346| 0.321| 0.371| 0.362|
| RND2     | 0.752| 0.695| 0.365| 0.346| 0.371| 0.375|     |
| RND3     | 0.721| 0.308| 0.289| 0.314| 0.305|     |     |
| RND4     | 0.384| 0.327| 0.390| 0.387|     |     |     |
| PRI1     | -    | -    | -    | 0.905| 0.924| 0.927|     |
| PRI2     | -    | -    | -    | -    | 0.911| 0.908|     |
| PRI3     | -    | -    | -    | -    | -    | 0.927|     |

(b) Q

| System   | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|----------|------|------|------|------|------|------|------|
| RND1     | 0.695| 0.730| 0.781| 0.387| 0.359| 0.394| 0.422|
| RND2     | 0.705| 0.654| 0.337| 0.327| 0.362| 0.378|     |
| RND3     | 0.733| 0.340| 0.292| 0.321| 0.349|     |     |
| RND4     | 0.422| 0.356| 0.403| 0.451|     |     |     |
| PRI1     | -    | -    | -    | 0.883| 0.879| 0.902|     |
| PRI2     | -    | -    | -    | -    | 0.889| 0.841|     |
| PRI3     | -    | -    | -    | -    | -    | 0.883|     |

(c) nERR

| System   | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|----------|------|------|------|------|------|------|------|
| RND1     | 0.571| 0.632| 0.695| 0.333| 0.302| 0.340| 0.337|
| RND2     | 0.679| 0.597| 0.298| 0.305| 0.330| 0.308|     |
| RND3     | 0.638| 0.257| 0.251| 0.270| 0.254|     |     |
| RND4     | 0.219| 0.187| 0.251| 0.241|     |     |     |
| PRI1     | -    | -    | -    | 0.886| 0.886| 0.895|     |
| PRI2     | -    | -    | -    | -    | 0.905| 0.908|     |
| PRI3     | -    | -    | -    | -    | -    | 0.908|     |

(d) iRBU

| System   | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|----------|------|------|------|------|------|------|------|
| RND1     | 0.616| 0.524| 0.495| 0.381| 0.343| 0.356| 0.365|
| RND2     | 0.597| 0.606| 0.390| 0.359| 0.371| 0.381|     |
| RND3     | 0.711| 0.508| 0.457| 0.483| 0.486|     |     |
| RND4     | 0.460| 0.448| 0.435| 0.476|     |     |     |
| PRI1     | -    | -    | -    | 0.848| 0.886| 0.902|     |
| PRI2     | -    | -    | -    | -    | 0.873| 0.889|     |
| PRI3     | -    | -    | -    | -    | -    | 0.883|     |

To visualise the general trend, $\tau$’s larger than 0.8 are shown in **bold**, and those smaller than 0.5 are shown in *italics*.

### 6.2 Why Does PRI Produce Similar Rankings?

This section discusses why different PRI-based qrels files produce similar rankings. To examine this phenomenon closely, Figure 7 visualises the relationship between the total number of **highly relevant** and **relevant** labels obtained and the document presentation order as seen by the assessors when WWW3E8 was constructed. The $x$-axis represents the document ranks shown on the PLY assessment interface. As the minimum pool size across the 160 topics was 128 (i.e., every topic had at least 128 pooled documents), we count the assessors’ labels (**highly relevant** or **relevant**) across the 160 topics for ranks 1–128. As each topic was judged by four assessors for each document ordering strategy, the maximum possible value for the $y$-axis is $4 * 160 = 640$: this
Table 9. Comparison of Mean System Ranking $\tau$’s Based on the $\tau$’s Shown in Table 8

| Measure | Mean $\tau$ (sample size) | Residual variance $V_{E1}$ | $p$-value (effect size) |
|---------|-----------------|-----------------|-----------------|
| nDCG   | RND-RND ($n_1 = 6$) | 0.741            | 0.000831 | $\approx 0$ (13.6) |
|        | PRI-PRI ($n_1 = 6$) | 0.917            | 0.00153  | $\approx 0$ (19.8) |
|        | RND-PRI ($n_1 = 16$) | 0.348            | 0.00172  | $\approx 0$ (13.0) |
| Q      | RND-RND ($n_1 = 6$) | 0.716            | 0.00270  | $\approx 0$ (14.9) |
|        | PRI-PRI ($n_1 = 6$) | 0.880            | 0.00172  | $\approx 0$ (14.9) |
|        | RND-PRI ($n_1 = 16$) | 0.369            | 0.00172  | $\approx 0$ (14.9) |
| nERR   | RND-RND ($n_1 = 6$) | 0.635            | 0.00172  | $\approx 0$ (14.9) |
|        | PRI-PRI ($n_1 = 6$) | 0.898            | 0.00172  | $\approx 0$ (14.9) |
|        | RND-PRI ($n_1 = 16$) | 0.280            | 0.00172  | $\approx 0$ (14.9) |
| iRBU   | RND-RND ($n_1 = 6$) | 0.592            | 0.00172  | $\approx 0$ (14.9) |
|        | PRI-PRI ($n_1 = 6$) | 0.880            | 0.00172  | $\approx 0$ (14.9) |
|        | RND-PRI ($n_1 = 16$) | 0.419            | 0.00172  | $\approx 0$ (14.9) |

Results of Tukey HSD tests for unpaired data with sample sizes 6, 6, and 16 are shown. The effect sizes are standardised mean differences based on the one-way ANOVA residual variance $V_{E1}$ [29].

would happen if, at a particular rank, all four assessors gave a highly relevant or relevant label for all 160 topics. Note that we are using the full topic set for this analysis as it does not involve computation of evaluation measures.

It is clear from Figure 7 that while the counts of highly relevant and relevant labels across topics are not correlated under the RND condition, we obtain more and more highly relevant and relevant labels as we approach the top of the PRI-based document ranks. There are two possible (mutually nonexclusive) explanations for this phenomenon: (I) the pseudorelevance as computed by NTCIRPOOL is accurate to some degree, and often manages to present truly relevant documents before nonrelevant ones; (II) under the PRI condition, the assessors tend to overrate the documents that they encounter early. Recall that, for each topic, each assessor receives either an RND pool or a PRI pool at random; they are not even aware that there are two kinds of document ordering strategies. The sharp contrast shown in Figure 7 despite this blind nature of the experiment suggests that the PRI strategy tends to prioritise documents that immediately look relevant (e.g., documents that contain query terms in the title field). That is, it is possible that because they look relevant, assessors tend to label them as so. Again, note that this does not rule out Explanation (I): the documents that look relevant may often be truly relevant.

Recall that the document sort keys for the PRI strategy are (a) the number of runs that returned the document; and (b) the sum of the ranks of that document in each of the above runs. In essence, the PRI strategy orders documents based on majority votes of the participating runs, and the assessors tend to agree with the majority votes. That is, these “popular” documents tend to be rated highly regardless of who the assessor is. Put another way, different PRI-based qrels files produce similar system rankings probably because they are all “similarly biased” towards popular documents.

7 RQ4: ROBUSTNESS TO NEW SYSTEMS

We now know that RND-based and PRI-based labels substantially differ from each other, and that the system ranking similarities under the PRI condition are higher than those under the RND condition. However, a more practically important question is RQ4: which strategy is more robust to the evaluation of systems that did not contribute to the pool? It is known that relevance assessments of test collections (especially those based on a small pool depth) are incomplete, and that new systems tend to be underrated if evaluated with such collections [25, 26, 32, 46, 47], because the new systems may return relevant documents that are outside the pools. While researchers should be aware of this, we still would not want test collections to fail catastrophically when evaluating new systems.

The robustness to new systems can be quantified using Leave-One-Team-Out (LOTO) tests [26, 32, 46]. That is, for each of the eight versions of qrels and for each team (G) that participated in the NTCIR-15 WWW-3 task [37], we remove G’s unique contributions from the original qrels.
Fig. 7. Document presentation order vs. counts of relevance labels (relevant, highly relevant) based on the main experiment with the 160 topics.

To form a “leave-out-G” qrels file, here, a unique contribution is a topicdoc that was originally contributed to the pool by team G and by no other teams. The WWW-3 task received runs from nine teams, and therefore we created $8 \times 9 = 72$ LOTO qrels files from the full qrels files covering the aforementioned 147 topics. Table 10 shows the relevant statistics of our LOTO experiments. For example, by removing the 1,612 unique contributions of Group 1 from the full qrels file that contained 29,522 topicdocs, we create a “leave-out-Group-1” qrels, with which we can simulate a situation where “new” runs from Group 1 are evaluated using an existing test collection. We then compare the system ranking based on the original qrels with the new ranking based on each LOTO qrels in terms of Kendall’s $\tau$. If the $\tau$ is low, that means that the LOTO qrels substantially underrate the “new” runs, which by extension suggests that the original qrels file is also not robust to real new runs that did not contribute to the pools.
Table 10. Number of Topicdocs in Leave-one-team-out Qrels Files

| team left out | #runs | unique contributions in LOTO qrels |
|--------------|------|------------------------------------|
| Group 1      | 3    | 1,612                              |
| Group 2      | 5    | 1,347                              |
| Group 3      | 5    | 1,351                              |
| Group 4      | 5    | 5,857                              |
| Group 5      | 5    | 3,651                              |
| Group 6      | 5    | 3,799                              |
| Group 7      | 5    | 858                                |
| Group 8      | 1    | 124                                |
| Group 9      | 3    | 468                                |

The original qrels size covering the 147 topics is 29,522.

Table 11. Mean System Ranking $\tau$ Over Nine Leave-one-team Out Experiments

|        | RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 | $V_{E2}$ |
|--------|------|------|------|------|------|------|------|------|----------|
| nDCG   | 0.914| 0.893| 0.901| 0.910| 0.956| 0.959| 0.957| 0.965| 0.000876 |
| Q      | 0.898| 0.876| 0.884| 0.902| 0.940| 0.947| 0.943| 0.959| 0.001340 |
| nERR   | 0.927| 0.905| 0.912| 0.922| 0.969| 0.973| 0.972| 0.975| 0.000600 |
| iRBU   | 0.923| 0.910| 0.907| 0.901| 0.965| 0.966| 0.957| 0.966| 0.000859 |

For example, the RND1 column compares the original RND1 qrels with nine leave-one-team-out versions of the qrels. For each evaluation measure, a paired Tukey HSD test at the 5% significance level was conducted. $V_{E2}$ denotes the two-way ANOVA residual variance for computing effect sizes.

Table 11 shows, for each of the eight qrels files (PRI1 through RND4) and for each evaluation measure, the mean $\tau$ scores averaged over the nine LOTO trials. Table 12 shows the accompanying results of the paired Tukey HSD tests. They can be summarised as follows:

- None of the differences within each document ordering strategy are statistically significant.
- All statistically significant differences are cases where a PRI-based qrels file outperforms a RND-based qrels file. The largest effect size observed is over 2.0 for every evaluation measure, e.g., 2.86 for nERR and 2.21 for iRBU as shown in Table 12(c) and (d).

Hence, the answer to RQ4 is clear: The PRI strategy substantially outperforms the RND strategy in terms of robustness to new systems. The result suggests that the PRI strategy often helps the assessors identify relevant documents that affect the evaluation of many systems, regardless of whether the systems contributed to the pools or not. Put another way, because the RND strategy ignores the “popularity” of documents, it is liable to miss relevant documents that are useful for evaluating many systems fairly.

To examine the above result more closely, Figures 8 and 9 visualise the LOTO results with RND2 and PRI4 for nDCG, whose mean $\tau$’s are the lowest and the highest among the eight versions of qrels (0.893 and 0.965 as shown in Table 11, respectively). The $y$-axis represents the mean nDCG scores, while the $x$-axis represents the runs from all nine groups sorted according to the full qrels file over the 147 topics. For example, “Group1-1” means Run 1 from Group 1. The results of leaving out Groups 8 and 9 are omitted as their unique contributions are small (See Table 10), and therefore the curves are very similar to the one for the full qrels. Runs that are heavily underrated by a
Table 12. All Statistically Significantly Different Pairs of Qrels Versions in Terms of Robustness to New Systems (Mean $\tau$ Over n = 9 Leave-one-team-out Experiments), Based on a Paired Tukey HSD Test at the 5% Significance Level

| qrels pairs     | p-value      | effect size | qrels pairs     | p-value      | effect size |
|-----------------|--------------|-------------|-----------------|--------------|-------------|
| (a) mean nDCG   |              |             | (b) mean Q      |              |             |
| PRI4-RND2       | 0.0000783    | 2.45        | PRI4-RND2       | 0.000303     | 2.27        |
| PRI2-RND2       | 0.000417     | 2.22        | PRI4-RND3       | 0.00142      | 2.05        |
| PRI4-RND3       | 0.000610     | 2.17        | PRI2-RND2       | 0.00288      | 1.95        |
| PRI3-RND2       | 0.000679     | 2.13        | PRI3-RND2       | 0.00645      | 1.83        |
| PRI1-RND2       | 0.000888     | 2.12        | PRI1-RND2       | 0.0111       | 1.74        |
| PRI2-RND3       | 0.00294      | 1.94        | PRI2-RND3       | 0.0117       | 1.73        |
| PRI3-RND3       | 0.00462      | 1.88        | PRI4-RND1       | 0.0173       | 1.67        |
| PRI4-RND4       | 0.00486      | 1.87        | PRI3-RND3       | 0.0244       | 1.61        |
| PRI1-RND3       | 0.00591      | 1.84        | PRI4-RND4       | 0.0329       | 1.56        |
| PRI4-RND1       | 0.0108       | 1.75        | PRI1-RND3       | 0.0398       | 1.53        |
| PRI2-RND4       | 0.0200       | 1.64        | PRI3-RND4       | 0.0298       | 1.58        |
| PRI3-RND4       | 0.0298       | 1.58        | PRI1-RND4       | 0.0369       | 1.54        |
| PRI1-RND4       | 0.0369       | 1.54        | PRI2-RND1       | 0.0410       | 1.52        |
| (c) mean nERR   |              |             | (d) mean iRBU   |              |             |
| PRI4-RND2       | 0.0000032    | 2.86        | PRI2-RND4       | 0.000445     | 2.21        |
| PRI2-RND2       | 0.0000059    | 2.78        | PRI4-RND4       | 0.000457     | 2.21        |
| PRI3-RND2       | 0.0000098    | 2.72        | PRI1-RND4       | 0.000496     | 2.20        |
| PRI1-RND2       | 0.0000198    | 2.63        | PRI2-RND3       | 0.00200      | 2.00        |
| PRI4-RND3       | 0.0000312    | 2.57        | PRI4-RND3       | 0.00205      | 2.00        |
| PRI2-RND3       | 0.0000564    | 2.49        | PRI1-RND3       | 0.00222      | 1.99        |
| PRI3-RND3       | 0.0000913    | 2.43        | PRI2-RND2       | 0.00352      | 1.92        |
| PRI1-RND3       | 0.000180     | 2.34        | PRI4-RND2       | 0.00361      | 1.91        |
| PRI4-RND4       | 0.000678     | 2.15        | PRI1-RND2       | 0.00390      | 1.90        |
| PRI2-RND4       | 0.00117      | 2.08        | PRI3-RND4       | 0.00410      | 1.90        |
| PRI3-RND4       | 0.00183      | 2.01        | PRI3-RND3       | 0.0159       | 1.68        |
| PRI4-RND1       | 0.00242      | 1.97        | PRI3-RND2       | 0.0261       | 1.60        |
| PRI1-RND4       | 0.00339      | 1.92        | PRI2-RND1       | 0.00406      | 1.90        |
| PRI2-RND1       | 0.00406      | 1.90        | PRI3-RND1       | 0.00617      | 1.83        |
| PRI1-RND1       | 0.0110       | 1.74        |

Effect sizes are based on the two-way ANOVA residual variances shown in Table 11.

LOTO qrels file can be identified as a “V” in the curves.\(^{22}\) For example, in Figure 9, it is easy to observe from the “lo-Group4” curve that if Group 4 is left out, this group’s runs (e.g., Group 4–5) are heavily underrated. Leaving out this particular group disrupts the ranking this much because this group had as many as 5,857 unique contributions to the pools (See Table 10). If we compare Figures 8 and 9, it can be observed that:

\(^{22}\) This visualisation approach for LOTO tests was used earlier by Sakai et al. [32] for the purpose of evaluating the robustness of diversified search evaluation measures.
Fig. 8. Effect of leaving one team out from the RND2 qrels (Mean nDCG over 147 topics).
Fig. 9. Effect of leaving one team out from the PRI4 qrels (Mean nDCG over 147 topics).
(i) Compared to PRI4, RND2 gives similar scores to all runs, suggesting that PRI is indeed biased towards "popular" documents.

(ii) For both RND2 and PRI4, the top half of the runs do not suffer much even when they are left out from the pools; it is the bottom half of the runs that are heavily underrated when treated as new runs. Moreover, the LOTO qrels files with RND2 suffer from this more often (i.e., there are more large "V"s).

Based on the above LOTO results that show the robustness of the PRI-based pools, if researchers need to evaluate their current IR systems using legacy IR test collections, we recommend the use of those constructed using the PRI approach unless they have a good reason to believe that their systems retrieve relevant documents that are vastly different from the pooled documents. Put another way, if the researchers believe that their new systems return search results that are reasonably similar to existing systems, then a PRI-based test collection is recommended; otherwise the new systems may be heavily underrated.

8 ADDITIONAL EXPERIMENTS: POOL DEPTH AND RUN QUALITY

This section reports on additional experiments that utilise subsets of the RND-based and PRI-based qrels files to re-examine RQ3 (system ranking agreement discussed in Section 6) and RQ4 (robustness to new systems discussed in Section 7). The additional experiments were conducted to address two specific questions from a reviewer: (I) What happens to the difference between RND and PRI if the pool depth is smaller? (II) What happens to the difference between RND and PRI if the submitted runs are less effective? We did not incorporate these additional results into Sections 6 and 7 because (a) the experimental settings of the additional experiments are somewhat artificial, which prevents us from making strong claims; (b) the results of these additional experiments are similar to our main experiments for RQ3 and RQ4 and therefore do not affect our main conclusions; and (c) we wanted to maintain the conciseness of Sections 6 and 7, which address our original research questions.

8.1 Method

Recall that our main experiments for RQ3 (system ranking agreement) and RQ4 (robustness to new systems) were based on depth-15 pools, with 29,522 topicdocs covering 147 topics (See Section 6.1). We address Reviewer Question (I) (effect of reducing the pool depth) as follows:

(1) Form a list of topicdocs by extracting all documents from each of the 37 submitted run files, where the document ranks are between 1 and 5.

(2) Obtain a subset of each RND-based and PRI-based qrels file, by using the above topicdoc list as a filter.

Thus, although our relevance labels are still from the depth-15-based experiments, the labelled documents we utilise in this additional experiments are those that qualify even if the pool depth was 5. Note that this only approximates the depth-5 situation because (a) the assessors processed the depth-15 pools in the presented order; and (b) for PRI-based pools, the pseudo relevance scores based on depth-15 pools and those based on depth-5 pools generally differ, since the number of runs that contain a document at or above rank 15 is generally not the same as the number of runs that contain that document at or above rank 5. Nevertheless, the above setting is probably a reasonable approximation. We shall refer to these filtered qrels files as rr1-5 versions, where rr stands for "ranks in the runs."

We addressed Reviewer Question (II) (effect of poorly performing runs) in a similar way: instead of extracting documents ranked between 1 and 5 in Step (1) discussed above, we extracted
documents ranked between 11 and 15 (i.e., documents presumed worst in the depth-15 pool files). This approximates a situation where the effectiveness of the runs are relatively poor and different runs return different documents, which should disrupt the majority-voting approach of PRI. We shall refer to these filtered qrels files as \texttt{rr11-15} versions. While this experiment is also a rough simulation of a real situation, note that by comparing the outcomes of our \texttt{rr11-15} and \texttt{rr1-5} experiments, we can discuss the effect of the run quality under the same condition, where each run contributes exactly five documents to the pool.

In the additional experiments reported below, evaluation measure scores are averaged over 137 topics for the following reasons. (1) As our \texttt{rr1-5} and \texttt{rr11-15} versions of the qrels files are subsets of the original ones, for some topics we ended up with no relevant documents. The union of such filtered topics across all 16 (i.e., 8 \texttt{rr1-5} and 8 \texttt{rr11-15}) qrels files amounts to 17 topics (0012,0024,0026,0027,0044,0060,0063,0132,0144,0147,0152,0153,0163,0169,0174,0177,0179). (2) Furthermore, in our LOTO experiments based on the above reduced versions of the qrels files, we lost 6 topics in addition (0014,0047,0058,0143,0150,0175). Hence, in order to evaluate the runs using a common topic set for every experimental condition, we average the evaluation measures over 160 − 17 − 6 = 137 topics.

The \texttt{rr1-5} qrel files each contains 11,989 topicdocs covering the 137 topics, whereas, the \texttt{rr11-15} qrels files each contains 15,089 topicdocs. The latter number is larger because the documents ranked between 11 and 15 by each run are indeed less similar than those ranked between 1 and 5 by each run. That is, the tendency with the \texttt{rr11-15} setting is indeed that different runs contribute different documents.

### 8.2 Revisiting RQ3: System Ranking Agreement

Table 13 shows the system ranking agreement results using the \texttt{rr1-5} (i.e., “depth-5”) versions of the qrels, and Table 14 shows the corresponding Tukey HSD test results for comparing the mean \(\tau\) within RND, within PRI, and between RND and PRI. These tables are arranged in the same way as Tables 8 and 9. Similarly, Table 15 shows the system ranking agreement results using the \texttt{rr11-15} (i.e., “poor runs”) versions of the qrels, and Table 16 shows the corresponding Tukey HSD test results.

It can be observed that the results under the \texttt{rr1-5} and \texttt{rr11-15} conditions are similar to our main results shown in Tables 8 and 9. That is, even under these conditions where each run contributes exactly five documents to the pool, the PRI-based qrels files produce very similar rankings, the RND-base qrels files produce moderately similar rankings, and the RND-based rankings and the PRI-based rankings are substantially different. Again, Tables 14 and 16 show that these differences are statistically highly significant. The effect sizes in the rightmost column show that the differences between the within-RND and within-PRI agreements are about one to five standard deviations apart, with the \texttt{rr11-15} (“poor quality runs”) results showing slightly smaller effects. This is not surprising as the \texttt{rr11-15} experiment is designed to disrupt the majority voting-based approach of PRI.

We concluded in Section 6.1 that different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition, and that PRI-based rankings and RND-based rankings are substantially different from each other. The additional experiments with the \texttt{rr1-5} and \texttt{rr11-15} versions of the qrels files suggest that these may still hold true even if the pool depth is as small as five, and even if the runs are less effective than those actually submitted to the NTCIR-15 WWW-3 task. The reader is reminded, however, that these results should be taken with a grain of salt for the reasons discussed in Section 8.1.
Table 13. System Ranking Agreement as Measured by Kendall’s $\tau$ between System Ranking Pairs ($n = 36$ Runs). According to the Four Official Measures of the NTCIR-15 WWW-3 Task, Based on the **rr1-5** Versions of the Qrels (Averaged Over 137 Topics)

(a) nDCG

|       | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|-------|------|------|------|------|------|------|------|
| RND1  | 0.854| 0.860| 0.844| 0.432| 0.422| 0.486| 0.460|
| RND2  | -    | 0.860| 0.813| 0.476| 0.467| 0.524| 0.498|
| RND3  | -    | -    | 0.851| 0.451| 0.435| 0.505| 0.460|
| RND4  | -    | -    | -    | 0.473| 0.438| 0.514| 0.489|
| PRI1  | -    | -    | -    | -    | 0.927| 0.927| 0.927|
| PRI2  | -    | -    | -    | -    | -    | 0.898| 0.905|
| PRI3  | -    | -    | -    | -    | -    | -    | 0.911|

(b) Q

|       | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|-------|------|------|------|------|------|------|------|
| RND1  | 0.841| 0.863| 0.870| 0.495| 0.492| 0.530| 0.508|
| RND2  | -    | 0.851| 0.794| 0.502| 0.505| 0.537| 0.502|
| RND3  | -    | -    | 0.860| 0.492| 0.489| 0.527| 0.492|
| RND4  | -    | -    | -    | 0.486| 0.476| 0.521| 0.498|
| PRI1  | -    | -    | -    | -    | 0.952| 0.927| 0.917|
| PRI2  | -    | -    | -    | -    | -    | 0.930| 0.895|
| PRI3  | -    | -    | -    | -    | -    | -    | 0.914|

(c) nERR

|       | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|-------|------|------|------|------|------|------|------|
| RND1  | 0.651| 0.641| 0.730| 0.311| 0.273| 0.324| 0.330|
| RND2  | -    | 0.629| 0.648| 0.362| 0.375| 0.413| 0.394|
| RND3  | -    | -    | 0.657| 0.302| 0.270| 0.289| 0.314|
| RND4  | -    | -    | -    | 0.238| 0.225| 0.276| 0.276|
| PRI1  | -    | -    | -    | -    | 0.873| 0.886| 0.898|
| PRI2  | -    | -    | -    | -    | -    | 0.898| 0.886|
| PRI3  | -    | -    | -    | -    | -    | -    | 0.892|

(d) iRBU

|       | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|-------|------|------|------|------|------|------|------|
| RND1  | 0.800| 0.721| 0.692| 0.381| 0.375| 0.425| 0.400|
| RND2  | -    | 0.756| 0.752| 0.441| 0.448| 0.467| 0.473|
| RND3  | -    | -    | 0.806| 0.508| 0.476| 0.533| 0.508|
| RND4  | -    | -    | -    | 0.543| 0.556| 0.562| 0.549|
| PRI1  | -    | -    | -    | -    | 0.860| 0.886| 0.917|
| PRI2  | -    | -    | -    | -    | -    | 0.886| 0.879|
| PRI3  | -    | -    | -    | -    | -    | -    | 0.917|

To visualise the general trend, $\tau$’s larger than 0.9 are shown in **bold**, and those smaller than 0.7 are shown in *italics*.

8.3 Revisiting RQ4: Robustness to New Systems

We also reran the LOTO experiments for evaluating the robustness of qrels files to new systems (i.e., **RQ4**), starting with the **rr1-5** ("depth-5") and **rr11-15** ("poor runs") qrels files described in Section 8.1. Tables 17–19 show the results under the **rr1-5** setting; they are arranged in exactly the same way as our main results (Tables 10–12). Similarly, Tables 20–22 show the results under the **rr11-15** setting. It can be observed that our new results are similar to the main results discussed in Section 7. That is, (1) None of the differences within each document ordering strategy are
Table 14. Comparison of Mean System Ranking $\tau$’s Based on the $\tau$’s Shown in Table 13

| Measure | Mean $\tau$ (sample size) | Residual variance $V_{E_1}$ | $p$-value (effect size) |
|---------|----------------------------|-----------------------------|------------------------|
|         | RND-RND $(n_1 = 6)$ | RND-PRI $(n_1 = 16)$ | RND-RND vs RND-PRI | PRI-PRI vs RND-PRI | PRI-PRI vs RND-RND |
| nDCG    | 0.847                  | 0.916                      | 0.471                  | 0.000659           | $\approx 0$ (14.7) | $\approx 0$ (17.3) | 0.000269 (2.68) |
| Q       | 0.847                  | 0.923                      | 0.503                  | 0.000404           | $\approx 0$ (17.1) | $\approx 0$ (20.9) | 0.000002 (3.78) |
| nERR    | 0.659                  | 0.889                      | 0.311                  | 0.002012           | $\approx 0$ (7.77) | $\approx 0$ (12.9) | 0.00002 (5.12)  |
| iRBU    | 0.755                  | 0.891                      | 0.478                  | 0.002833           | $\approx 0$ (5.20) | $\approx 0$ (7.76) | 0.000457 (2.56) |

Results of Tukey HSD tests for unpaired data with sample sizes 6, 6, and 16 are shown. The effect sizes are standardised mean differences based on the one-way ANOVA residual variance $V_{E_1}$ [29].

Table 15. System Ranking Agreement as Measured by Kendall’s $\tau$ between System Ranking Pairs ($n = 36$ Runs). According to the Four Official Measures of the NTCIR-15 WWW-3 Task, Based on the rr11-15 Versions of the Qrels (Averaged Over 137 Topics)

| Measure | RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|---------|------|------|------|------|------|------|------|------|
| (a) nDCG | 0.870 | 0.870 | 0.895 | 0.632 | 0.610 | 0.619 | 0.619 | 0.603 |
| RND2    | -    | 0.860 | 0.886 | 0.603 | 0.581 | 0.590 | 0.590 | 0.562 |
| RND3    | -    | -    | 0.886 | 0.622 | 0.594 | 0.603 | 0.603 | 0.581 |
| RND4    | -    | -    | -    | 0.580 | 0.568 | 0.578 | 0.578 | 0.556 |
| PRI1    | -    | -    | -    | -    | 0.927 | 0.949 | 0.949 | 0.927 |
| PRI2    | -    | -    | -    | -    | -    | 0.914 | 0.914 | 0.911 |
| PRI3    | -    | -    | -    | -    | -    | -    | 0.911 | 0.911 |
| PRI4    | -    | -    | -    | -    | -    | -    | -    | 0.911 |

(b) Q

| RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|------|------|------|------|------|------|------|------|
| 0.860 | 0.844 | 0.873 | 0.695 | 0.644 | 0.679 | 0.679 | 0.651 |
| 0.863 | 0.873 | 0.657 | 0.619 | 0.629 | 0.629 | 0.629 | 0.619 |
| 0.895 | 0.868 | 0.629 | 0.663 | 0.663 | 0.648 | 0.648 | 0.648 |
| 0.638 | 0.575 | 0.629 | 0.629 | 0.600 | 0.600 | 0.600 | 0.600 |
| PRI1 | -    | -    | -    | -    | 0.930 | 0.895 | 0.895 | 0.895 |
| PRI2 | -    | -    | -    | -    | -    | -    | 0.911 | 0.911 |
| PRI3 | -    | -    | -    | -    | -    | -    | -    | 0.911 |

(c) nERR

| RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|------|------|------|------|------|------|------|------|
| 0.838 | 0.825 | 0.870 | 0.575 | 0.514 | 0.587 | 0.587 | 0.543 |
| 0.822 | 0.816 | 0.616 | 0.568 | 0.622 | 0.622 | 0.622 | 0.578 |
| PRI1 | -    | -    | -    | -    | 0.902 | 0.911 | 0.911 | 0.911 |
| PRI2 | -    | -    | -    | -    | -    | 0.889 | 0.889 | 0.889 |
| PRI3 | -    | -    | -    | -    | -    | -    | -    | 0.889 |

(d) iRBU

| RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 |
|------|------|------|------|------|------|------|------|
| 0.841 | 0.800 | 0.784 | 0.524 | 0.533 | 0.581 | 0.581 | 0.508 |
| 0.813 | 0.797 | 0.587 | 0.590 | 0.625 | 0.625 | 0.625 | 0.584 |
| 0.870 | 0.635 | 0.638 | 0.686 | 0.613 | 0.613 | 0.613 | 0.613 |
| PRI1 | -    | -    | -    | -    | 0.889 | 0.905 | 0.905 | 0.921 |
| PRI2 | -    | -    | -    | -    | -    | 0.895 | 0.895 | 0.895 |
| PRI3 | -    | -    | -    | -    | -    | -    | 0.883 | 0.883 |

To visualise the general trend, $\tau$’s larger than 0.9 are shown in bold, and those smaller than 0.7 are shown in italics.
Table 16. Comparison of Mean System Ranking \( \tau \)'s Based on the \( \tau \)'s Shown in Table 15

| Measure | RND-RND (\( n_1 = 6 \)) | PRI-PRI (\( n_1 = 6 \)) | RND-PRI (\( n_1 = 16 \)) | Residual variance \( V_{E1} \) | RND-RND vs PRI-PRI | PRI-PRI vs PRI-PRI | PRI-PRI vs RND-RND |
|---------|--------------------------|--------------------------|--------------------------|-------------------------------|-------------------|-------------------|-------------------|
| nDCG    | 0.878                    | 0.926                    | 0.593                    | 0.000355                      | \( \approx 0 \) (15.1) | \( \approx 0 \) (17.6) | 0.000488 (2.55)   |
| Q       | 0.868                    | 0.913                    | 0.641                    | 0.000737                      | \( \approx 0 \) (8.35) | \( \approx 0 \) (10.0) | 0.0232 (1.64)     |
| nERR    | 0.831                    | 0.904                    | 0.556                    | 0.001194                      | \( \approx 0 \) (7.98) | \( \approx 0 \) (10.1) | 0.00327 (2.11)    |
| iRBU    | 0.818                    | 0.898                    | 0.604                    | 0.001801                      | \( \approx 0 \) (5.02) | \( \approx 0 \) (6.91) | 0.00858 (1.89)    |

Results of Tukey HSD tests for unpaired data with sample sizes 6, 6, and 16 are shown. The effect sizes are standardised mean differences based on the one-way ANOVA residual variance \( V_{E1} \) [29].

Table 17. Number of Topicdocs in Leave-one-team-out Qrels Files Under the rr1-5 Setting ("depth-5")

| team left out | #runs | unique contributions | #topicdocs in LOTO qrels |
|---------------|-------|----------------------|--------------------------|
| Group 1       | 3     | 714                  | 11,275                   |
| Group 2       | 5     | 565                  | 11,424                   |
| Group 3       | 5     | 1,019                | 10,970                   |
| Group 4       | 5     | 2,237                | 9,752                    |
| Group 5       | 5     | 1,503                | 10,486                   |
| Group 6       | 5     | 1,551                | 10,438                   |
| Group 7       | 5     | 570                  | 11,419                   |
| Group 8       | 1     | 112                  | 11,877                   |
| Group 9       | 3     | 456                  | 11,533                   |

The original qrels size covering the 137 topics is 11,989.

Table 18. Mean System Ranking \( \tau \) Over Nine Leave-one-team Out Experiments Under the rr1-5 Setting ("depth-5")

|                         | RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 | \( V_{E2} \) |
|-------------------------|------|------|------|------|------|------|------|------|-------------|
| nDCG                    | 0.895| 0.878| 0.874| 0.890| 0.925| 0.932| 0.927| 0.942| 0.000885    |
| Q                       | 0.883| 0.862| 0.871| 0.883| 0.896| 0.911| 0.918| 0.922| 0.001171    |
| nERR                    | 0.871| 0.850| 0.858| 0.871| 0.935| 0.939| 0.941| 0.942| 0.001205    |
| iRBU                    | 0.892| 0.891| 0.877| 0.878| 0.941| 0.934| 0.944| 0.940| 0.000894    |

For example, the RND1 column compares the original RND1 qrels with nine leave-one-team-out versions of the qrels. For each evaluation measure, a paired Tukey HSD test at the 5% significance level was conducted. \( V_{E2} \) denotes the two-way ANOVA residual variance for computing effect sizes.

Statistically significant; and (2) All statistically significant differences are cases where a PRI-based qrels file outperforms a RND-based qrels file.

In Section 7, we concluded that the PRI strategy substantially outperforms the RND strategy in terms of robustness to new systems, and that the PRI strategy often helps the assessors identify relevant documents that affect the evaluation of many systems, regardless of whether the systems contributed to the pools or not. Our rr1-5 and rr11-15 results suggest that these conclusions may generalise to cases where the pool depth is as small as five, and even if the runs are not very effective. Again, however, the reader is reminded that these experiments only approximate real situations with depth-5 pools and poorly performing runs.
Table 19. All Statistically Significantly Different Pairs of Qrels Versions (Under the $rr1$-$5$) Setting in Terms of Robustness to New Systems (Mean $\tau$ Over $n = 9$ Leave-one-team-out Experiments), Based on a Paired Tukey HSD Test at the 5% Significance Level

| qrels pairs   | $p$-value | effect size | qrels pairs   | $p$-value | effect size |
|---------------|-----------|-------------|---------------|-----------|-------------|
| (a) mean nDCG |           |             | (b) mean $Q$  |           |             |
| PRI4-RND3     | 0.000316  | 2.26        | PRI4-RND2     | 0.00925   | 1.77        |
| PRI4-RND2     | 0.000751  | 2.14        | PRI3-RND2     | 0.0202    | 1.64        |
| PRI2-RND3     | 0.00283   | 1.95        | PRI4-RND3     | 0.0434    | 1.51        |
| PRI2-RND2     | 0.00626   | 1.83        |               |           |             |
| PRI3-RND3     | 0.00963   | 1.76        |               |           |             |
| PRI4-RND4     | 0.0125    | 1.72        |               |           |             |
| PRI1-RND3     | 0.0143    | 1.70        |               |           |             |
| PRI3-RND2     | 0.0201    | 1.64        |               |           |             |
| PRI1-RND2     | 0.0292    | 1.58        |               |           |             |
| PRI4-RND1     | 0.0292    | 1.58        |               |           |             |
| (c) mean nERR |           |             | (d) mean iRBU |           |             |
| PRI4-RND2     | 0.0000153 | 2.66        | PRI3-RND3     | 0.000363  | 2.24        |
| PRI3-RND2     | 0.0000222 | 2.61        | PRI3-RND4     | 0.000559  | 2.18        |
| PRI2-RND2     | 0.0000298 | 2.57        | PRI1-RND3     | 0.000749  | 2.14        |
| PRI1-RND2     | 0.0000792 | 2.45        | PRI4-RND3     | 0.000855  | 2.12        |
| PRI4-RND3     | 0.0000852 | 2.44        | PRI1-RND4     | 0.00114   | 2.08        |
| PRI3-RND3     | 0.000122  | 2.39        | PRI4-RND4     | 0.00130   | 2.06        |
| PRI2-RND3     | 0.000163  | 2.35        | PRI2-RND3     | 0.00343   | 1.92        |
| PRI1-RND3     | 0.000419  | 2.22        | PRI2-RND4     | 0.00509   | 1.86        |
| PRI4-RND1     | 0.00131   | 2.06        | PRI3-RND2     | 0.0107    | 1.75        |
| PRI4-RND4     | 0.00150   | 2.04        | PRI3-RND1     | 0.0135    | 1.71        |
| PRI3-RND1     | 0.00183   | 2.01        | PRI1-RND2     | 0.0198    | 1.65        |
| PRI3-RND4     | 0.00209   | 1.99        | PRI4-RND2     | 0.0221    | 1.63        |
| PRI2-RND1     | 0.00239   | 1.98        | PRI1-RND1     | 0.0246    | 1.61        |
| PRI2-RND4     | 0.00272   | 1.96        | PRI4-RND1     | 0.0274    | 1.59        |
| PRI1-RND1     | 0.00562   | 1.85        |               |           |             |
| PRI1-RND4     | 0.00637   | 1.83        |               |           |             |

Effect sizes are based on the two-way ANOVA residual variances shown in Table 11.

9 CONCLUSIONS

The present study addressed a few questions that remained open for the past two decades or so regarding two document ordering strategies for relevance assessors: PRI (practiced at NTCIR) and RND (recommended elsewhere, e.g., at TREC). Our experiments, which involved eight independent relevance assessments for 32,375 topic-document pairs (i.e., a total of 259,000 labels), provide some answers to them. Our conclusions are as follows:

RQ1: Which strategy enables more efficient relevance assessments? The choice of document ordering strategy has negligible impact on assessor efficiency statistics such as average time between judging two documents.

RQ2: Which strategy enables higher inter-assessor agreements? While the choice of the strategy substantially affects which documents are judged (highly) relevant, the difference between
Table 20. Number of Topicdocs in Leave-one-team-out Qrels Files Under the \texttt{rr11-15} Setting ("Low Run Quality")

| team left out | #runs | unique contributions | #topicdocs in LOTO qrels |
|---------------|-------|----------------------|--------------------------|
| Group 1       | 3     | 900                  | 14,189                   |
| Group 2       | 5     | 1,046                | 14,043                   |
| Group 3       | 5     | 1,076                | 14,013                   |
| Group 4       | 5     | 2,529                | 12,560                   |
| Group 5       | 5     | 1,838                | 13,251                   |
| Group 6       | 5     | 1,928                | 13,161                   |
| Group 7       | 5     | 907                  | 14,182                   |
| Group 8       | 1     | 120                  | 14,969                   |
| Group 9       | 3     | 568                  | 14,521                   |

The original qrels size covering the 137 topics is 15,089.

Table 21. Mean System Ranking \( \tau \) Over Nine Leave-one-team Out Experiments Under the \texttt{rr11-15} Setting ("Low Run Quality")

|        | RND1 | RND2 | RND3 | RND4 | PRI1 | PRI2 | PRI3 | PRI4 | \( V_{E2} \) |
|--------|------|------|------|------|------|------|------|------|------------|
| nDCG   | 0.949| 0.944| 0.955| 0.966| 0.978| 0.977| 0.975| 0.980| 0.000259   |
| Q      | 0.950| 0.947| 0.955| 0.959| 0.970| 0.965| 0.957| 0.974| 0.000323   |
| nERR   | 0.938| 0.940| 0.949| 0.946| 0.970| 0.974| 0.980| 0.972| 0.000235   |
| iRBU   | 0.957| 0.950| 0.952| 0.951| 0.975| 0.967| 0.980| 0.975| 0.000201   |

For example, the RND1 column compares the original RND1 qrels with nine leave-one-team-out versions of the qrels. For each evaluation measure, a paired Tukey HSD test at the 5\% significance level was conducted. \( V_{E2} \) denotes the two-way ANOVA residual variance for computing effect sizes.

The inter-assessor agreement under the PRI condition and that under the RND condition is of no practical significance.

**RQ3:** Which strategy enables more stable system rankings across different versions of qrels files? Different system rankings under the PRI condition are substantially more similar to one another than those under the RND condition. Moreover, PRI-based rankings and RND-based rankings are substantially different from each other.

**RQ4:** Which strategy is more robust to systems that did not contribute to the pools? Our leave-one-team-out results show that the PRI-based relevance assessments are substantially more robust. Our interpretation of this finding is that, while the PRI-based qrels files are probably biased towards popular documents, PRI often helps the assessors identify relevant documents that affect the evaluation of many systems, including those that did not contribute to the pools.

Furthermore, our somewhat artificial additional experiments suggest that our findings regarding **RQ3** and **RQ4** may hold true even if the pool depth is smaller than 15 and even if the runs are less effective than they are in our experiments.

From our answer to **RQ4**, if researchers need to evaluate their current IR systems using legacy IR test collections, we recommend the use of those constructed using the PRI approach unless they have a good reason to believe that their systems retrieve relevant documents that are vastly different from the pooled documents. Put another way, if the researchers believe that their new systems return search results that are reasonably similar to existing systems, then a PRI-based test collection is recommended; otherwise the new systems may be heavily underrated. While this
Table 22. All Statistically Significantly Different Pairs of Qrels Versions (Under the \textit{rr11-15} Setting in Terms of Robustness to New Systems (Mean $\tau$ Over $n = 9$ Leave-one-team-out Experiments), Based on a Paired Tukey HSD Test at the 5% Significance Level

| qrels pairs | $p$-value | effect size | qrels pairs | $p$-value | effect size |
|-------------|-----------|-------------|-------------|-----------|-------------|
| (a) mean nDCG | | | (b) mean Q | | | |
| PRI4-RND2   | 0.000377  | 2.24        | PRI4-RND2   | 0.0470    | 1.50        |
| PRI1-RND2   | 0.000967  | 2.11        | PRI1-RND2   | 0.00139   | 2.05        |
| PRI2-RND2   | 0.00150   | 2.04        | PRI2-RND2   | 0.00329   | 1.93        |
| PRI4-RND1   | 0.00319   | 1.93        | PRI4-RND1   | 0.00329   | 1.93        |
| PRI3-RND2   | 0.00384   | 1.91        | PRI3-RND2   | 0.00329   | 1.93        |
| PRI1-RND1   | 0.00756   | 1.80        | PRI1-RND1   | 0.00756   | 1.80        |
| PRI2-RND1   | 0.0112    | 1.74        | PRI2-RND1   | 0.0112    | 1.74        |
| PRI3-RND1   | 0.0258    | 1.60        | PRI3-RND1   | 0.0258    | 1.60        |
| PRI4-RND3   | 0.0355    | 1.55        | PRI4-RND3   | 0.0355    | 1.55        |
| (c) mean nERR | | | (d) mean iRBU | | | |
| PRI3-RND1   | 0.0000104 | 2.71        | PRI3-RND2   | 0.00118   | 2.08        |
| PRI3-RND2   | 0.0000256 | 2.59        | PRI3-RND4   | 0.00139   | 2.05        |
| PRI2-RND1   | 0.000206  | 2.32        | PRI3-RND3   | 0.00329   | 1.93        |
| PRI3-RND4   | 0.000458  | 2.21        | PRI4-RND2   | 0.0129    | 1.72        |
| PRI2-RND2   | 0.000482  | 2.20        | PRI1-RND2   | 0.0136    | 1.71        |
| PRI4-RND1   | 0.000626  | 2.17        | PRI4-RND4   | 0.0149    | 1.69        |
| PRI4-RND2   | 0.00142   | 2.05        | PRI1-RND4   | 0.0157    | 1.68        |
| PRI1-RND1   | 0.00129   | 2.06        | PRI3-RND1   | 0.0218    | 1.63        |
| PRI3-RND3   | 0.00202   | 2.00        | PRI4-RND3   | 0.0314    | 1.57        |
| PRI1-RND2   | 0.00286   | 1.95        | PRI1-RND3   | 0.0329    | 1.56        |
| PRI2-RND4   | 0.00678   | 1.82        | PRI2-RND4   | 0.0176    | 1.67        |
| PRI4-RND4   | 0.0176    | 1.67        | PRI2-RND3   | 0.0248    | 1.61        |
| PRI1-RND4   | 0.0319    | 1.56        | PRI1-RND4   | 0.0319    | 1.56        |

Effect sizes are based on the two-way ANOVA residual variances shown in Table 11.
REFERENCES

[1] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. 2008. Million query track 2007 overview.

[2] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen Voorhees. 2018. TREC common core track overview. In Proceedings of the TREC 2017.

[3] Peter Bailey, Nick Craswell, Ian Soboroff, Paul Thomas, Arjen P. de Vries, and Emine Yilmaz. 2008. Relevance assessment: Are judges exchangeable and does it matter?. In Proceedings of the ACM SIGIR 2008. 667–674.

[4] Ben Carterette, Virgil Pavlu, Hui Fang, and Evangelos Kanoulas. 2010. Million query track 2009 overview. In Proceedings of the TREC 2009.

[5] Alexandra Chouldechova and David Mease. 2013. Differences in search engine evaluations between query owners and non-owners. In Proceedings of the ACM WSDM 2013. 103–112.

[6] Charles L. A. Clarke, Nick Craswell, and Ian Soboroff. 2010. Overview of the TREC 2009 web track. In Proceedings of the TREC 2009.

[7] Cyril Cleverdon and Michael Keen. 1966. Factors Determining the Performance of Indexing Systems; Volume 2. Technical Report. College of Aeronautics, Cranfield, UK.

[8] Cyril Cleverdon, Jack Mills, and Michael Keen. 1966. Factors Determining the Performance of Indexing Systems; Volume 1: Design. Technical Report. College of Aeronautics, Cranfield, UK.

[9] Gordon V. Cormack, Christopher R. Palmer, and Charles L. A. Clarke. 1998. Efficient construction of large test collections. In Proceedings of the ACM SIGIR’98. 282–289.

[10] Tadele T. Damesse, J. Shane Culpepper, Jaewon Kim, and Falk Scholer. 2018. Presentation ordering effects on assessor agreement. In Proceedings of the ACM CIKM 2018. 723–732.

[11] Tadele T. Damesse, Falk Scholer, Kalervo Järvelin, and J. Shane Culpepper. 2016. The effect of document order and topic difficulty on assessor agreement. In Proceedings of the ACM ICTIR 2016. 73–76.

[12] Michael Eisenberg and Carol Barry. 1988. Order effects: A study of the possible influence of presentation order on user judgments of document relevance. Journal of the American Society for Information Science 39, 5 (1988), 293–300.

[13] Nicola Ferro and Carol Peters (Eds.), 2019. Information Retrieval in a Changing World: Lessons Learned from 20 Years of CLEF. Springer.

[14] Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. Psychological Bulletin 76, 5 (1971), 378–382.

[15] Donna K. Harman. 2005. The TREC test collections. In Proceedings of the TREC: Experiment and Evaluation in Information Retrieval, Ellen M. Voorhees and Donna K. Harman (Eds.), The MIT Press, Chapter 2.

[16] Mu-Hsuan Huang and Hui-Yu Wang. 2004. The influence of document presentation order and number of documents judged on users’ judgments of relevance. Journal of the American Society for Information Science 55, 11 (2004), 970–979.

[17] Noriko Kando. 2004. Evaluation of information access technologies at the NTCIR workshop. In Proceedings of the CLEF 2003 (LNCS 3237). 29–43.

[18] Klaus Krippendorff. 2018. Content Analysis: An Introduction to Its Methodology (Fourth Edition). SAGE Publications.

[19] Aldo Lipani, David E. Losada, Guido Zuccon, and Mihai Lupu. 2021. Fixed-cost pooling strategies. IEEE Transactions on Knowledge and Data Engineering 33, 4 (2021), 1503–1522.

[20] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2017. Multi-armed bandits for ordering judgements in pooling-based evaluation. Information Processing and Management 53, 3 (2017), 1005–1025.

[21] David E. Losada, Javier Parapar, and Álvaro Barreiro. 2018. When to stop making relevance judgments? A study of stopping methods for building information retrieval test collections. Journal of the Association for Information Science and Technology 70, 3 (2018), 49–60.

[22] Cheng Luo, Tetsuya Sakai, Yiqun Liu, Zhicheng Dou, Chenyan Xiong, and Jingfang Xu. 2017. Overview of the NTCIR-13 we want web task. In Proceedings of the NTCIR-13. 394–401.

[23] Jiaxin Mao, Tetsuya Sakai, Cheng Luo, Peng Xiao, Yiqun Liu, and Zhicheng Dou. 2019. Overview of the NTCIR-14 we want web task. In Proceedings of the NTCIR-14. 455–467.

[24] Masaki Muraoka, Zhaohao Zeng, and Tetsuya Sakai. 2020. SLWWW at the NTCIR-15 WWW-3 task. In Proceedings of the NTCIR-15. 243–246.

[25] Tetsuya Sakai. 2008. Comparing metrics across TREC and NTCIR: The robustness to system bias. In Proceedings of the ACM CIKM 2008. 581–590.

[26] Tetsuya Sakai. 2014. Metrics, statistics, tests. In Proceedings of the PROMISE Winter School 2013: Bridging between Information Retrieval and Databases (LNCS 8173). 116–163.

[27] Tetsuya Sakai. 2016. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006–2015. In Proceedings of the ACM SIGIR 2016. 5–14.

[28] Tetsuya Sakai. 2016. Topic set size design. Information Retrieval Journal 19, 3 (2016), 256–283.
[29] Tetsuya Sakai. 2018. Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power. Springer.

[30] Tetsuya Sakai. 2019. How to run an evaluation task. In Proceedings of the Information Retrieval Evaluation in a Changing World, Nicola Ferro and Carol Peters (Eds.). Springer.

[31] Tetsuya Sakai. 2021. On the instability of diminishing return IR measures. In Proceedings of the ECIR 2021 Part I (LNCS 12656). 572–586.

[32] Tetsuya Sakai, Zhicheng Dou, Ruihua Song, and Noriko Kando. 2012. The reusability of a diversified search test collection. In Proceedings of the AIRS 2012 (LNCS 7675). 26–38.

[33] Tetsuya Sakai, Noriko Kando, Chuan-Jie Lin, Teruko Mitamura, Hideki Shima, Donghong Ji, Kuang-Hua Chen, and Eric Nyberg. 2008. Overview of the NTCIR-7 ACLIA IRAQA task. In Proceedings of the NTCIR-7. 77–114.

[34] Tetsuya Sakai and Chin-Yew Lin. 2010. Ranking retrieval systems without relevance assessments - revisited. In Proceedings of the EVA 2010. 25–33.

[35] Tetsuya Sakai, Douglas W. Oard, and Noriko Kando (Eds.). 2020. Evaluating Information Retrieval and Access Tasks: NTCIR’s Legacy of Research Impact. Springer.

[36] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2021. WWW3E8: 259,000 relevance labels for studying the effect of document presentation order for relevance assessors. In Proceedings of the ACM SIGIR 2021. to appear.

[37] Tetsuya Sakai, Sijie Tao, Zhaohao Zeng, Yukun Zheng, Jiaxin Mao, Zhumin Chu, Yiqun Liu, Zhicheng Dou, Nicola Ferro, Maria Maistro, and Ian Soboroff. 2020. Overview of the NTCIR-15 we want web with CENTRE (WWW-3) task. In Proceedings of the NTCIR-15. 219–234.

[38] Tetsuya Sakai and Peng Xiao. 2019. Randomised vs. prioritised pools for relevance assessments: Sample size considerations. In Proceedings of the AIRS 2019 (LNCS 12004). 94–105.

[39] Tetsuya Sakai and Zhaohao Zeng. 2019. Which diversity evaluation measures are “Good”? In Proceedings of the ACM SIGIR 2019. 595–604.

[40] Tetsuya Sakai and Zhaohao Zeng. 2020. Retrieval evaluation measures that agree with users’ SERP preferences: Traditional, preference-based, and diversity measures. ACM TOIS 39, 2 Article 14 (2020). https://dl.acm.org/doi/10.1145/3431813.

[41] Falk Scholer, Diane Kelly, Wan-Ching Wu, Hanseul S. Lee, and William Webber. 2013. The effect of threshold priming and need for cognition on relevance calibration and assessment. In Proceedings of the ACM SIGIR 2013. 623–632.

[42] K. Sparck Jones and R. G. Bates. 1977. Report on a Design Study for the “Ideal” Information Retrieval Test Collection. Technical Report. Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5481.

[43] K. Sparck Jones and C. J. van Rijsbergen. 1975. Report on the Need for and Provision of an “Ideal” Information Retrieval Test Collection. Technical Report. Computer Laboratory, University of Cambridge, British Library Research and Development Report No. 5266.

[44] Andrew Turpin, Falk Scholer, Stefano Mizzaro, and Eddy Maddalena. 2015. The benefits of magnitude estimation relevance assessments for information retrieval evaluation. In Proceedings of the ACM SIGIR 2015. 565–574.

[45] Ellen M. Voorhees. 2000. Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing and Management 36, 5 (2000), 697–716.

[46] Ellen M. Voorhees. 2002. The philosophy of information retrieval evaluation. In Proceedings of the CLEF 2001 (LNCS 2406). 355–370.

[47] Justin Zobel. 1998. How reliable are the results of large-scale information retrieval experiments?. In Proceedings of the ACM SIGIR’98. 307–314.

Received May 2021; revised August 2021; accepted October 2021