On the Ordering of Pooled Web Pages, Gold Assessments, and Bronze Assessments

TETSUYA SAKAI, SIJIE TAO, NUO CHEN, and YUJING LI, Waseda University, Japan
MARIA MAISTRO, University of Copenhagen, Denmark
ZHUMIN CHU, Tsinghua University, P. R. C.
NICOLA FERRO, University of Padua, Italy

The present study leverages a recent opportunity we had to create a new English web search test collection for the NTCIR-16 We Want Web (WWW-4) task, which concluded in June 2022. More specifically, through the test collection construction effort, we examined two factors that may affect the relevance assessments of depth-k pools, which in turn may affect the relative evaluation of different IR systems. The first factor is the document ordering strategy for the assessors, namely, prioritisation (PRI) and randomisation (RND). PRI is a method that has been used in NTCIR tasks for over a decade; it ranks the pooled documents by a kind of pseudorelevance for the assessors. The second factor is assessor type, i.e., Gold or Bronze. Gold assessors are the topic creators and therefore they "know" which documents are (highly) relevant and which are not; Bronze assessors are not the topic creators and may lack sufficient knowledge about the topics. We believe that our study is unique in that the authors of this paper served as the Gold assessors when creating the WWW-4 test collection, which enabled us to closely examine why Bronze assessments differ from the Gold ones. Our research questions examine assessor efficiency (**RQ1**), inter-assessor agreement (**RQ2**), system ranking similarity with different qrels files (**RQ3**), system ranking robustness to the choice of test topics (**RQ4**), and the reasons why Bronze assessors tend to be more liberal than Gold assessors (**RQ5**). The most remarkable of our results are as follows. Firstly, in the comparisons for **RQ1** through **RQ4**, it turned out that what may matter more than the document ordering strategy (PRI vs. RND) and the assessor type (Gold vs. Bronze) is how well-motivated and/or well-trained the Bronze assessors are. Secondly, regarding **RQ5**, of the documents originally judged nonrelevant by the Gold assessors contrary to the Bronze assessors in our experiments, almost one half were truly relevant according to the Gold assessors' own reconsiderations. This result suggests that even Gold assessors are far from perfect; budget permitting, it may be beneficial to hire highly-motivated Bronze assessors in addition to Gold assessors so that they can complement each other.

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

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

1 INTRODUCTION

Decades after the proposal of 'ideal' test collections in the 1970s [46, 47], offline evaluation of IR systems using pooling-based test collections is still a vital tool for advancing the state-of-the-art while ensuring reproducibility in the IR community. The present study concerns the *presentation order* of pooled documents for relevance

Authors' addresses: Tetsuya Sakai, tetsuyasakai@acm.org; Sijie Tao, tsjmailbox@ruri.waseda.jp; Nuo Chen, pleviumtan@toki.waseda.jp; Yujing Li, liyujing@ruri.waseda.jp, Waseda University, 3-4-1 Okubo, Shinjuku, Tokyo, Japan, 169-8555; Maria Maistro, mm@di.ku.dk, University of Copenhagen, Universitetsparken 1, Copenhagen, Denmark, 2100; Zhumin Chu, chuzm19@mails.tsinghua.edu.cn, Tsinghua University, Haidian District, Beijing, P. R. C., 100084; Nicola Ferro, ferro@dei.unipd.it, University of Padua, Via Giovanni Gradenigo 6/b, Padova, Italy, I-35131.

1046-8188/2023/5-ART $15.00
https://doi.org/10.1145/3600227

assessors as well as the *assessor type* [5] under a *depth-k pooling* [18, 34] setting. It follows up on a large-scale study of Sakai et al. [40, 41],[1] which compared the following two document ordering strategies.

Randomisation (RND) presents the pooled documents in random order to remove the *rank bias* of the assessors: that is, to prevent the assessors from overrating the highly ranked documents and underrating the later documents.

Prioritisation (PRI) sorts the pooled documents by a simple pseudorelevance score using the NTCIRPOOL script.[2] The first sort key is the number of runs that returned the document at or above depth k (larger the better); the second sort key is the sum of the ranks of the document within those runs (smaller the better).

The PRI method reflects the view that "popular" (i.e., "liked by many systems") documents are likely to be relevant. From the very beginning of NTCIR, documents pooled at NTCIR were sorted for the assessors using a strategy similar to PRI [21]; In 2008, Sakai et al. [37] introduced the PRI method for an NTCIR-7 task, which later became available in NTCIRPOOL; this tool has been used in many NTCIR tasks since then.

The main findings of Sakai et al. [41] were that assessors tend to label "popular" documents as relevant and this very bias towards popular documents may make the test collection more robust to the handling of new systems. This is because the popular documents affect the evaluation of many systems, including those that did not contribute to the pools. However, like other studies that considered multiple document ordering strategies for assessors (See Section 3.3), their entire study was conducted under a *Bronze* assessor environment [5]: their assessors were *not* the owners of the queries; nor were they topic experts.

This study attempts to generalise the Bronze-based work of Sakai et al. [41] by considering in addition *Gold* assessors, i.e., topic creators [5], for the following reasons. In an ideal world, we would like to always hire Gold assessors to construct qrels (query-relevance sets) files as by definition these people are the ones with information needs and therefore should be the ones to determine what is relevant and what is not. Nevertheless, due to practical difficulties, we usually make do with Bronze assessors, *assuming* that they are reasonable substitutes. Despite this common practice, however, several studies have suggested that the above assumption does not always hold (See Section 2.1). Moreover, despite the above definition of the Gold assessors, they are human and may make mistakes; how exactly they behave relative to Bronze assessors under the PRI and RND environments has never been explored before. Gaining insight into different combinations of assessor type and document ordering strategy will help track/task organisers make design decisions: for example, *if* there is some substantial benefit in hiring Gold assessors, they can consider actually doing so (as we have done at NTCIR) instead of hiring only Bronze assessors, even if this incurs an additional cost.

The main research questions of the present study are as follows.

RQ1 (assessor efficiency) **RQ1.1**: Which document ordering strategy (i.e., PRI or RND) enables more *efficient* relevance assessments for Gold assessors? **RQ1.2**: Which assessor type (i.e., Gold or Bronze) enables more *efficient* relevance assessments under a PRI environment?

RQ2 (inter-assessor agreement) How do the document ordering strategy and assessor type affect inter-assessor agreement?

RQ3 (system ranking similarity) **RQ3.1**: How *similar* are PRI-based and RND-based system rankings under a Gold environment? **RQ3.2**: How *similar* are Gold-based and Bronze-based system rankings under a PRI environment?

[1] Note that Sakai et al. [41] is the corrected version of Sakai et al. [40] after a major bug fix. A corrigendum to Sakai et al. [40] is available at https://dl.acm.org/action/downloadSupplement?doi=10.1145%2F3494833&file=p76-sakai-corrigendum.pdf, which points to Sakai et al. [41]. The present study was conducted after the bug fix.

[2] http://research.nii.ac.jp/ntcir/tools/ntcirpool-en.html

RQ4 (system ranking consistency) **RQ4.1**: How *robust to the choice of test data* are PRI-based and RND-based system rankings under a Gold environment? **RQ4.2**: How *robust to the choice of test data* are Gold-based and Bronze-based system rankings under a PRI environment?

RQ5 (liberal Bronze assessors) Why do Bronze assessors judge *more documents* to be relevant compared to Gold assessors?

Note that the above research questions do not involve Gold-Bronze comparisons under a RND environment. This is because PRI is the widely used document ordering strategy at NTCIR, and accordingly all of our Bronze assessments rely on PRI pool files.

The present study is unique both as a study on assessor type (i.e., Gold vs. Bronze) and as a study on document ordering strategies (i.e., PRI vs. RND) in that every author of this paper served as a Gold assessor when creating a new English web search test collection for NTCIR-16 [38]. Unlike prior art in which researchers commented on Gold-Bronze disagreements as a third party, our arrangement enables us to directly address questions such as **RQ5**, as *we* are the right answers by definition (except when we make human errors). On the other hand, we acknowledge that our study is smaller in scale compared to the Bronze-based experiments of Sakai et al. [41], as we shall discuss in Section 2.3. We also acknowledge that we made no special attempt at making our Gold and Bronze assessors behave homogeneously within each assessor group beyond giving them the same environment with the same instructions; hence different assessors might have been motivated to different degrees. However, we shall demonstrate in Section 4.3 that they are in fact reasonably homogeneous within each assessor groups and therefore that our comparisons across groups are meaningful.

The remainder of this paper is organised as follows. Section 2 discusses related work, and Section 3 describes our data, which we constructed through our effort of running the NTCIR-16 We Want Web (WWW-4) Task [38]. Sections 4-8 address **RQ1**-**RQ5**, respectively. Finally, Section 9 concludes this paper. In addition, the Appendix discusses an additional research question (**RQ6**) regarding the robustness of relevance assessments to new systems: we refrain from including it in the main body of this paper, as our sample size for this particular experiment (i.e., the number of participating teams, each of which is left out in turn) is too small to obtain conclusive results, unlike the Bronze-based experiments of Sakai et al. [41].

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; it is not possible to discuss them exhaustively in this paper. Below, we focus our attention on existing studies on the effect of assessor type (Section 2.1), on pooling and document ordering strategies (Section 2.2), and on the work of Sakai et al. [41], which compared the PRI and RND strategies only under the Bronze environment (Section 2.3).

2.1 Gold, Silver, and Bronze Assessors

Bailey et al. [5] defined the following three types of assessors in 2008. *Gold* (topic originators), *Silver* (task experts who are not topic originators), and *Bronze* (who are neither topic originators nor task experts). Hereafter, we often refer to this taxonomy to provide a unified view of prior art, even when discussing work that predates their work.

The Cranfield II test collection relied on Gold assessors ("questioners"), and later hired three people (two from the Aircraft Research Association and one from the College of Aeronautics: hence probably Silver assessors) to form alternative relevance assessments. Cleverdon [9] reported that the system ranking was robust to the switching of the relevance assessment sets. Three decades later, in 2000, Voorhees [50] obtained similar conclusions based on TREC-4 and TREC-6 data (with Bronze assessments from a TREC participant for the TREC-6 test collection [12]).

In 2012, Alonso and Mizzaro [4] argued that combining multiple crowd workers' assessments (Bronze) can be a good substitute for TREC assessments (Gold); the study was based on rejudging 206 topic-document pairs from TREC-7 and TREC-8, and the effect on system ranking was not discussed. Compared to these studies, some of the findings discussed below seem less optimistic.

In 2002, Sormunen [45] had 38 topics from TREC-7 and TREC-8 rejudged with graded relevance and found that the original (i.e., Gold) binary relevance assessments of TREC are quite *liberal*: many marginally relevant documents (according to their Bronze assessors) were considered relevant (by the Gold assessors).[3] We remark that at NTCIR, the situation is quite different from these early TRECs: NTCIR has used graded relevance since its launch [35], and we shall demonstrate in Section 3.4 that our graded Bronze assessments tend to be more liberal than our graded Gold assessments.

Using the TREC 2007 Enterprise track data, Bailey et al. [5], in 2008, concluded that Bronze assessors may not be a good substitute for Gold assessors due to "*unfamiliarity with task and topic context*." Also in 2008, Kinney et al. [23] sampled web search queries and compared the relevance assessments of domain experts (Silver) with those of "generalists" (Bronze), and reported that the shallow and inaccurate Bronze labels can be improved by providing "intent statements" written by experts. In 2012, Clough et al. [10] compared the assessments of one UK Government's National Archives employee (Silver) with those of crowd workers (Bronze), and the results were generally in line with the above studies.

In 2013, Chouldechova and Mease [7] reported on experiments where Google query owners (Gold) and non-owners (Bronze) were compared by means of side-by-side SERP preference tests rather than document relevance assessments. They reported that query owners were more reliable at choosing the better SERP and enable higher statistical power for this task. In 2014, Al-Harbi and Smucker [1] reported on a think-aloud study of student assessors using four TREC 2005 Robust Track topics. They present four categories of primary-secondary (i.e., Gold-Bronze) assessor disagreements: *difficulty in applying the search topic*, *difficulty in processing the document*, *assessor factors* (lack of knowledge or lack of concentration), and *error in the primary assessor's judgment*.

On a more optimistic note, Wakeling et al. [55] reported in 2016 that the inter-assessor agreements between primary (Gold) and secondary (Bronze) assessors were high in their user study which used Google's search results. However, as this was a user study designed to leverage real information needs, the number of Gold assessments was only 600, or 15 documents per topic.

Also in 2016, McDonnell et al. [29] compared the crowd workers' assessments (Bronze) with the official assessments of the TREC 2009 Web Track data, and found that making the crowd workers explicitly enter document excerpts as *judgement rationales* improves their accuracy.[4] This study was similar in scale to Wakeling et al. [55]: only 700 topic-document pairs from the TREC track were rejudged by crowd workers (with five workers for each document).[5] Hence, in a follow-up study in 2018, Kutlu et al. [24] compared crowd assessments (with rationales) with those of the official TREC 2014 Web track assessments using 4,991 topic-document pairs: they reported that the crowd assessments can provide a system ranking that is highly similar to the official ranking. Moreover, the authors analysed the cases where NIST and crowd assessors disagree, and presented a taxonomy: the top categories are *NIST error*, *crowd error*, *different perception of relevance*, *ambiguous topic definition*, and *technical issues*.

In order to closely examine Gold-Bronze disagreements, every author of this paper served as a Gold assessor when constructing the NTCIR-16 We Want Web (WWW-4) test collection as task organisers [38]. One particularly

[3] The TREC-8 overview paper states as follows. "*Ad hoc topics have been constructed by the same person who performed the relevance assessments for that topic (called the* assessor*) since TREC-3*" [54].

[4] Alonso and Mizzaro [4] let crowd workers enter free-text assessment "justifications."

[5] It may also be worth noting that the TREC 2009 Web track topics were developed based on a search engine query log, and also on a query clustering algorithm for the purpose of mining subtopics [8], and therefore that they are probably not real information needs of the NIST assessors, even if they are the topic creators.

unique aspect of our study is that we, as Gold assessors (not a third party as in prior art), *rejudged* the documents judged relevant only by the Bronze assessors to address **RQ5**. Through this additional effort, we did find some relevant documents that we missed at the test collection construction phase. Also, it is worth noting that, unlike some of the above small-scale studies, we constructed a full large-scale English web search test collection with 10,333 topic-document pairs, or 206.7 documents per topic, where each topic-document pair was judged by two Bronze assessors (from two different sites) in addition to one Gold assessor.

2.2 Pooling and Document Ordering for Assessors

Through experiments with the Bing search engine and crowd assessors (Bronze), Shokouhi et al. [44] reported that the assessors tend to assign different labels depending on the relevance of the previously labeled document. It then follows that document ordering for relevance assessors will impact the outcome of the assessments. Sakai et al. [40, Section 2] provides an overview of previous work concerning pooling and document ordering for relevance assessors, in the context of their PRI-RND comparison with Bronze assessors. Hence the following discussion overlaps considerably with theirs, although we provide comments from the assessor type viewpoint in addition.

After the launch of TREC in 1992 [18], alternatives to depth-k pooling were proposed (e.g., [2, 6, 14, 25, 58]), and a few such methods have been adopted by TREC tracks. For example, the TREC 2017 Common Core Track [3] adopted a version of the *MaxMean* method of Losada et al. [27], which dynamically selects which run to process based on the judgements so far. Subsequently, Voorhees [51] reported that the TREC 2017 Common Core test collection is not as *fair* (i.e., among the participating runs, some are more favoured over others than they should be) and *reusable* (i.e., new runs that did not contribute to the pools can be assessed appropriately relative to those that did) as desired. We observe that the Common Core Track relevance assessors are not Gold assessors since the topics were updated versions of the TREC 2004 Robust Track topics. More recently, the TREC 2021 Deep Learning test collection [53] was constructed using a dynamic document selection approach called *Continuous Active Learning* [13] with the MS MARCO v2 corpus.[6] Since the Deep Learning track topics are a small subsample of the MS MARCO queries, it follows that the Deep Learning track assessors are also not Gold assessors.

While acknowledging the benefits of these dynamic document selection approaches, the present study adheres to depth-k pooling because it is still widely used for its own advantages: compared to the dynamic strategies, it facilitates assessment cost estimation, and easily enables the assessors to *rejudge* documents (i.e., modify the relevance labels that they previously chose) [41]. Note that even the aforementioned Common Core Track relied on depth-10 pooling initially to accommodate a *burn-in period* for the assessors [3].

Regarding document ordering for assessors, early studies relied on small-scale experiments (with only *one* search topic) where printed documents were provided to the (Bronze) assessors [16, 19]. While these studies reported on a *hedging phenomenon* (i.e., the assessors were reluctant to label early documents with very high or very low scores, because they might want to reserve these extreme scores for later documents), Sakai et al. [40] observe that this may be largely due to their 7-point scale relevance ratings. As in Sakai et al. [41], our relevance assessors choose from *highly relevant*, *relevant*, or *nonrelevant* (and *error*; see Section 3.4) for each pooled document, which clearly is a simpler task. Moreover, Sakai et al. [40] point out that the above early studies *assume* that the relevance assessments obtained under a RND environment are the ground truth.

In 2013, Scholer et al. [43] studied the effect of the overall relevance of early documents on the 4-point (Bronze) 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 observed an effect similar to the hedging phenomenon of Eisenberg and Barry [16], and argued that "*people's*

[6] https://microsoft.github.io/msmarco/

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, Damessie et al. [15] compared PRI (using `NTCIRPOOL` as in our study), RND, and a third document ordering strategy, and their results suggested that PRI achieves a higher inter-assessor agreement than RND. However, their experiments relied on only 240 topic-document pairs: eight topics (4 from TREC-7, 4 from TREC-8), each with 30 pooled documents. Also in 2018, Losada et al. [28] reported on a simulation-based study on when to stop judging documents to reduce the assessment cost, under the premise that pooled documents are ranked by a kind of pseudorelevance.

None of the above studies on document ordering involved Gold assessors. To the best of our knowledge, the present study is the first to examine the effect of document ordering on Gold assessors.

2.3 Sakai et al. on Bronze Assessors

By constructing a new test collection for the NTCIR-16 WWW-4 task with both Gold and Bronze assessments, the present study complements the purely Bronze-based work of Sakai et al. [41].[7] Hence, this section summarises their previous findings with Bronze assessors.

Below, the main findings from Sakai et al. [41] are duplicated and labelled as "**B***n*", where the B stands for Bronze.

B1 (efficiency) There is no substantial difference between RND and PRI in terms of time spent for judging each document, although PRI may enable faster identification of the first highly relevant document in the pool.

B2 (inter-assessor agreement) The difference between the inter-assessor agreement under the RND condition and that under the PRI condition is probably of no practical significance.

B3 (system ranking similarity) While PRI-based qrels files tend to generate system ranking that are slightly more similar to each other than RND-based qrels files do, this difference is probably of no practical significance. On the other hand, system ranking similarities tend to be lower for PRI-RND comparisons than for PRI-PRI and RND-RND comparisons. The PRI strategy tends to make the assessor favour "popular" documents, i.e., those returned at high ranks by many systems.

B4 (robustness to new systems) PRI-based qrels files tend to be slightly more robust to new systems than RND-based ones. This is probably because the PRI strategy tends to help us identify "popular" relevant documents. The "popular" relevant documents affect the evaluation of many systems, including systems that did not contribute to the pools.

Our **RQ1**, **RQ2**, **RQ3**, **RQ6** (see the Appendix) correspond to **B1**, **B2**, **B3**, **B4**, respectively; The differences are that our new study involves Gold assessors in addition to Bronze assessors, with the new NTCIR-16 WWW-4 test collection [38]. Our **RQ4** (system ranking consistency) and **RQ5** (liberal Bronze assessors) are entirely new contributions.

We acknowledge, on the other hand, that the Bronze-based experiments of Sakai et al. [41] were larger in scale than ours from several different viewpoints: they had 160 topics (we only have 50); they had four PRI qrels files along with four RND files (we only have three qrels files in total: one Gold and two Bronze files); they had 36 runs from nine teams (we have 18 runs from four teams). However, the two studies are comparable in terms of the number of documents per topic: they had $(32,375/160 =)202.3$ documents per topic based on depth-15 pools; we have $(10,333/50 =)206.7$ documents per topic based on depth-60 pools.

[7] Sakai and Xiao [42] also reported on a preliminary study on the effect of PRI and RND using data from NTCIR-13 WWW-1 and NTCIR-14 WWW-2. However, their results also suffer from the same bug that is described in Sakai et al. [41].

3 DATA

The authors of this paper were part of the NTCIR-16 WWW-4 task organiser team [38]. Addressing our research questions was part of our plan when we started constructing the WWW-4 test collection, a new English web search test collection. This section describes how the WWW-4 test collection with both Gold and Bronze relevance assessments was constructed; hence it has a substantial overlap with the (unrefereed) WWW-4 overview paper [38],[8] but provides additional details, including how each assessor was assigned to a PRI-based or RND-based pool file of each topic.

3.1 Target Web Corpus: Chuweb21

The WWW-4 task introduced a new English web corpus called Chuweb21, which was constructed based on the April 2021 block of Common Crawl dataset.[9] Details of the corpus construction process can be found in the WWW-4 overview paper [38]. Chuweb21 contains 82, 451, 337 HTMLs or 1.69 TiB of compressed content; it is publicly available.[10]

3.2 Topics

The WWW-4 organisers determined the number of topics to create using Sakai's topic set size design tool for comparing $m = 2$ systems with ANOVA (or equivalently, with a t-test) [32].[11] Accordingly, they created 50 topics: according to the (corrected) WWW-3 results, this is expected to ensure a statistical power of more than 80% for a *minimum detectable difference* of 0.1 in nERR (normalised Expected Reciprocal Rank) and 0.05 in iRBU (intentwise Rank-Biased Utility) [39]. That is, whenever there is a true mean difference of at least 0.1 in terms of nERR (or at least 0.05 in terms of iRBU), it is expected that this can actually be detected as a statistically significant difference (at the 5% significance level) 80% of the time.[12]

The WWW-4 topic set is publicly available.[13] As can be seen in the XML file, each topic has a *content* field (e.g., "Timnit Gebru Google") and a *description* field (e.g., "I want to know the details regarding Google's firing of Dr. Timnit Gebru."). The authors of this paper (Sakai, Li, Ferro, Chen, Chu, Maistro, Tao, hereafter referred to as AssessorG1-AssessorG7, where the "G" stands for Gold) developed the topics by performing pilot searches on the Chuweb21 corpus using a search interface provided by Ian Soboroff (one of the WWW-4 organisers) making sure that we have at least one relevant document for each topic. All relevant documents found by the Gold assessors at this topic development step were recorded to form a run called ORG-TOPICDEV (See Section 3.3). As shown in the "Gold" column of Table 1, AssessorG1 created the first eight topics, and the other Gold assessors each created seven topics. These topics reflect the actual information needs and interests of each Gold assessor.

3.3 Constructing PRI and RND Pool Files

The WWW-4 task received a total of 18 runs from four participating teams including the organisers's team. University of Tsukuba, Waseda University, and Tsinghua University contributed 6, 5, and 5 runs, respectively. The organisers contributed a baseline BM25 run, and a run called ORG-TOPICDEV, which is simply a collection of relevant documents identified by the Gold assessors at the topic development stage. Because of the limited number of participating teams, we decided to form depth-60 pools for the relevance assessments to alleviate the relevance assessment *incompleteness* problem [30, 57]. As was mentioned earlier, this gave us a total of 10,333

[8] The WWW-4 overview paper also suffers from the aforementioned bug. Corrected results can be found in Sakai et al. [39].

[9] https://commoncrawl.org/2021/04/april-2021-crawl-archive-now-available/

[10] https://drive.google.com/drive/folders/11hi_R6cSIHEZx3QwyG5KQjgRVmxXhWta?usp=sharing

[11] http://www.f.waseda.jp/tetsuya/samplesizeANOVA2.xlsx

[12] For the WWW-3 data, nERR was the least stable measure (with a residual variance of $V_{E2} = 0.0284$) and iRBU was the most stable measure (with a residual variance of $V_{E2} = 0.00716$) [32].

[13] https://waseda.box.com/www4topicsxml A subset of this topic set is shown later in this paper, in Table 13.

Table 1. Assigned topics and pool files for the Gold and Bronze assessors. All Bronze assessors used the PRI-based pool file for each topic. The pool depth is 60 and the numbers of the second column add up to 10,333.

TopicID	#docs	Gold	Pool type for Gold	BronzeW (Waseda)	BronzeT (Tsinghua)
0201	140	AssessorG1	PRI	AssessorW5	AssessorT3
0202	210	AssessorG1	RND	AssessorW1	AssessorT4
0203	232	AssessorG1	RND	AssessorW2	AssessorT3
0204	215	AssessorG1	PRI	AssessorW4	AssessorT5
0205	211	AssessorG1	PRI	AssessorW4	AssessorT3
0206	276	AssessorG1	RND	AssessorW2	AssessorT1
0207	247	AssessorG1	RND	AssessorW4	AssessorT2
0208	223	AssessorG1	PRI	AssessorW5	AssessorT1
0209	199	AssessorG2	PRI	AssessorW5	AssessorT3
0210	184	AssessorG2	RND	AssessorW3	AssessorT4
0211	180	AssessorG2	RND	AssessorW1	AssessorT5
0212	162	AssessorG2	PRI	AssessorW1	AssessorT5
0213	173	AssessorG2	PRI	AssessorW3	AssessorT5
0214	157	AssessorG2	PRI	AssessorW4	AssessorT2
0215	197	AssessorG2	RND	AssessorW2	AssessorT2
0216	176	AssessorG3	RND	AssessorW5	AssessorT2
0217	266	AssessorG3	PRI	AssessorW1	AssessorT4
0218	179	AssessorG3	PRI	AssessorW3	AssessorT1
0219	305	AssessorG3	PRI	AssessorW2	AssessorT3
0220	244	AssessorG3	RND	AssessorW4	AssessorT4
0221	211	AssessorG3	RND	AssessorW5	AssessorT4
0222	179	AssessorG3	PRI	AssessorW1	AssessorT2
0223	203	AssessorG4	PRI	AssessorW3	AssessorT5
0224	187	AssessorG4	RND	AssessorW1	AssessorT1
0225	163	AssessorG4	PRI	AssessorW3	AssessorT3
0226	164	AssessorG4	RND	AssessorW3	AssessorT4
0227	179	AssessorG4	RND	AssessorW3	AssessorT1
0228	254	AssessorG4	PRI	AssessorW2	AssessorT2
0229	249	AssessorG4	PRI	AssessorW5	AssessorT2
0230	227	AssessorG5	RND	AssessorW1	AssessorT3
0231	210	AssessorG5	PRI	AssessorW5	AssessorT4
0232	221	AssessorG5	RND	AssessorW5	AssessorT3
0233	235	AssessorG5	RND	AssessorW2	AssessorT2
0234	272	AssessorG5	RND	AssessorW1	AssessorT4
0235	210	AssessorG5	PRI	AssessorW2	AssessorT1
0236	183	AssessorG5	PRI	AssessorW1	AssessorT5
0237	213	AssessorG6	RND	AssessorW4	AssessorT3
0238	171	AssessorG6	RND	AssessorW4	AssessorT2
0239	171	AssessorG6	PRI	AssessorW2	AssessorT5
0240	184	AssessorG6	RND	AssessorW4	AssessorT5
0241	237	AssessorG6	RND	AssessorW2	AssessorT4
0242	168	AssessorG6	PRI	AssessorW5	AssessorT1
0243	159	AssessorG6	RND	AssessorW3	AssessorT3
0244	185	AssessorG7	RND	AssessorW3	AssessorT5
0245	201	AssessorG7	PRI	AssessorW1	AssessorT1
0246	238	AssessorG7	PRI	AssessorW5	AssessorT4
0247	255	AssessorG7	RND	AssessorW2	AssessorT1
0248	196	AssessorG7	PRI	AssessorW3	AssessorT1
0249	274	AssessorG7	PRI	AssessorW4	AssessorT2
0250	158	AssessorG7	RND	AssessorW4	AssessorT5

topic-document pairs to judge, or 206.7 documents per topic on average. The exact pool size per topic is shown in the “#docs” column of Table 1.

For each topic, two pool files were created from the depth-60 pooled documents, using the PRI and RND strategies (See Section 1). Following the previous practices of NTCIR, all Bronze assessors (details to be given in Section 3.4) conducted their relevance assessments using the PRI files. On the other hand, for Gold assessors, we randomly assigned either a RND file or a PRI file for each topic, as shown in the “Pool type for Gold” column of

Table 1, while balancing the number of RND files and that of PRI files to process for each assessor. This was for addressing our research questions concerning Gold assessors.

3.4 Assessors and Qrels Files

The Gold assessors were not told whether they were given a RND file or a PRI file for each topic. As for Bronze assessors, we hired two groups of assessors independently at Waseda University, Japan ("BronzeW" assessors), and at Tsinghua University, China ("BronzeT" assessors). Each Bronze group had five assessors (AssessorW1-AssessorW5 and AssessorT1-AssessorT5), and each Bronze assessor handled 10 topics assigned at random, as shown in the "Bronze" columns of Table 1. The BronzeW assessor group comprised four master students and one undergraduate student from the English-based programme of the computer science department at Waseda University; The BronzeT assessors were professional labellers from a Chinese vendor, who are proficient in English.

All assessors used the browser-based PLY interface [34, 41] for conducting relevance assessments; the interface features a document list panel on the left, and a document content panel on the right, and the topic content panel at the top. Both the content and description fields of each topic (See Section 3.2) were displayed to the assessors. Hence, note that the Bronze assessors saw the Gold assessor's descriptions (e.g., "I want to know the details regarding Google's firing of Dr. Timnit Gebru.") in addition to the "queries" (e.g., "Timnit Gebru Google"). For each document, the assessor chose from *highly relevant*, *relevant*, *nonrelvant*, and *error* (the document is not displayed properly due to a problem in the HTML file etc.), which were later mapped to the relevance levels of L2, L1, L0, and L0, and the gain values of 2, 1, 0, and 0 for computing the official evaluation measures of the task. These are: nDCG (Microsoft version [31] of normalised Discounted Cumulative Gain [20]), Q-measure [31, 33], and the aforementioned nERR and iRBU, all measured at the document cutoff of 10. Once a document has been judged, the PLY interface leads to assessor to the next document in the document list, and therefore the assessor usually processes the document list from top to bottom. However, they can also select a particular document to modify their previous assessments; this behaviour will be discussed quantitatively in Section 4.

The assessors conducted their relevance assessments between December 21, 2021 and January 20, 2022. User activities on the PLY interface was recorded with timestamps at the backend and the present study utilises this data for addressing **RQ1** (assessor efficiency). Table 2 shows the relevance label statistics that we have obtained: we shall utilise the labels to address **RQ2**-**RQ5**. It can be observed that, on the whole, the Bronze assessors are a little more *liberal* than the Gold assessors, in the sense that they identified fewer L0 (i.e., nonrelevant) documents and more L2 (i.e., highly relevant) documents. As shown in the table, hereafter we refer to the combination of a PRI file and a Gold assessor as PRI-Gold assessments, and the combination of a RND file and a Gold assessor as RND-Gold assessments. Also, we shall refer to the topics that have PRI-Gold assessments as PRI-Gold topics, and those that have RND-Gold assessments as RND-Gold topics. Furthermore, to make explicit the fact that all Bronze assessors used a PRI pool for every topic, we shall refer to the Bronze assessments as PRI-BronzeW and PRI-BronzeT assessments. That is, each PRI-Gold topic has PRI-Gold, PRI-BronzeW, and PRI-BronzeT assessments, while each RND-Gold topic has RND-Gold, PRI-BronzeW, and PRI-BronzeT assessments.

4 RQ1: GOLD ASSESSOR EFFICIENCY

This section addresses **RQ1** (assessor efficiency). Following Sakai et al. [41], we analyse the following statistics.

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.

Table 2. Distribution of relevance labels over the three relevance levels (L2: highly relevant, L1: relevant; L0: nonrelevant).

Relevance level	Gold	(PRI-Gold/RND-Gold)	PRI-BronzeW	PRI-BronzeT
L2	1,373	(674/699)	1,591	1,776
L1	1,806	(918/888)	3,158	1,986
L0	7,154	(3,537/3,617)	5,584	6,571
Total	10,333	(5,129/5,204)	10,333	10,333

Table 3. Gold assessor efficiency results. Each n denotes the sample size per group. For each criterion, the result of a two-sample t-test is shown, and a $*$ indicates a statistical significance at $\alpha = 0.05$. The effect sizes (Glass's Δ) are computed using the standard deviation (s.d.) of the RND statistics.

Criterion	Mean PRI-Gold	Mean RND-Gold	p-value	s.d. (RND)	Glass's Δ
TJ1D (seconds)	41.7 ($n = 21$)	65.4 ($n = 25$)	$p = 0.0531$	47.5	0.499
TF1RH (seconds)	79.9 ($n = 24$)	120.7 ($n = 22$)	$p = 0.0364*$	58.2	0.702
TF1H (seconds)	127.7 ($n = 23$)	210.8 ($n = 20$)	$p = 0.0499*$	154.3	0.538
ATBJ (seconds)	25.7 ($n = 25$)	27.4 ($n = 25$)	$p = 0.664$	15.0	0.110
NREJ (times)	8.04 ($n = 25$)	3.44 ($n = 25$)	$p = 0.102$	4.66	0.986

Among these, we consider ATBJ to be the primary efficiency criterion, as it is an estimate of the time spent in judging one document, which can be used directly to estimate the total assessment cost in advance.

Prior to the analysis, we removed potential outliers from the samples as follows. For TJ1D and ATBJ, we removed instances that exceed 3 minutes, as we cannot tell from the activity log whether the assessor is working or has left the desk [41]. Moreover, for TF1RH and TF1H, we applied a cap of 10 minutes instead because these statistics generally represent time to read multiple documents until a (highly) relevant document is found. As discussed below, these caps may be quite rigorous, but they still give us sample sizes large enough to quantify the trends from a statistical point of view.

Section 4.1 addresses **RQ1.1** (*Which document ordering strategy enables more efficient relevance assessments for Gold assessors?*); Section 4.2 addresses **RQ1.2** (*Which assessor type enables more efficient relevance assessments under a PRI environment?*). In addition, Section 4.3 examines the assessor efficiency statistics at the *assessor* level rather than the *assessor-type* level. This is to validate the basic assumption behind our study, namely, that the assessors within each assessor type share similar characteristics; in this section we specifically examine the per-assessor *efficiency* statistics.

4.1 RQ1.1: PRI vs. RND with Gold Assessors (Efficiency)

Table 3 addresses **RQ1.1** (*Which document ordering strategy enables more efficient relevance assessments for Gold assessors?*) by comparing the Gold assessor efficiency statistics across the 25 PRI-Gold and the 25 RND-Gold topics (minus a few topics that were removed as described above). For each efficiency criterion, a statistical significance test result based on a two-sample t-test is shown, with and effect size in terms of Glass's Δ [17, 32]. For example, for TJ1D, Δ is computed from the table simply as $(65.4 - 41.7)/47.5 = 0.499$. The following observations can be made.

- The mean ATBJ for PRI and that for RND are similar and the difference is not statistically significant; the effect size is very small ($\Delta = 0.110$). This result is consistent with the ATBJ results of Sakai et al. [41] where

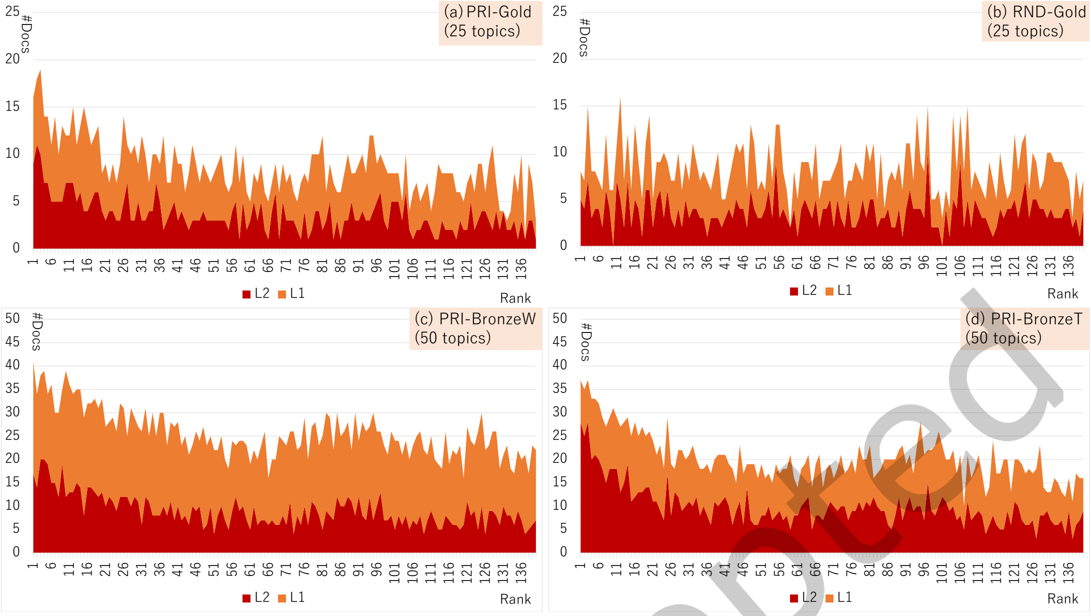

Fig. 1. Number of documents judged (highly) relevant against document presentation order. (a) and (b) show the number of Gold-relevant documents summed across the 25 PRI-Gold and 25 RND-Gold topics, respectively; (c) and (d) show the number of Bronze-relevant documents summed across all 50 topics.

only *Bronze* assessors were involved. Hence, we conclude that *there is no substantial difference between PRI and RND in terms of time spent for judging each document, even for Gold assessors.*

- The differences in means in terms of TF1RH and that in terms of TF1H are statistically significant (with Δ over 0.5). That is, *Gold assessors tend to identify the first (highly) relevant document more quickly under the PRI environment.* This is generally in line with the Bronze results of Sakai et al. [41]. Later in this section, we shall discuss why PRI enables faster identification of the first (highly) relevant document for a given topic.
- The difference in means in terms of TJ1D suggests that PRI may enable faster judging of the very first document, but it is not quite statistically significant. This weak trend is also consistent with the Bronze results of Sakai et al. [41].
- The difference in means in terms NREJ is not statistically significant either. While the effect size for NREJ is relatively large (Δ = 0.986), suggesting that Gold assessors tend to correct labels more often under the PRI environment, the reverse trend (also statistically not significant) was observed in the Bronze results of Sakai et al. [40, 41]. Hence we refraining from drawing a conclusion regarding NREJ.

In summary, *under the Gold environment, there is no substantial difference between PRI and RND in terms of time spent for judging each document, although assessors tend to find the first (highly) relevant document more quickly with PRI files.* These results for Gold assessors are generally in line with those for Bronze assessors [41, Table 2] (See also Section 2.3 **B1**).

Following Sakai et al. [41] where only Bronze assessments were examined (with data from the WWW-3 task), Figure 1 plots, for each of the PRI-Gold, RND-Gold, PRI-BronzeW, and PRI-BronzeT environments, the number of documents judged highly relevant (L2) and relevant (L1) against the document presentation order on the

Table 4. Gold and Bronze assessor efficiency results under the PRI environment. Each n denotes the sample size per group. For each criterion, the result of a unpaired Tukey HSD test is shown whenever the p-value is less than $\alpha = 0.05$. The leftmost column shows the residual variance V_{E1} of one-way ANOVA, for computing effect sizes [32].

Criterion	(a) Mean PRI-Gold	(b) Mean PRI-BronzeW	(c) Mean PRI-BronzeT	V_{E1}
TJ1D (seconds)	41.7 ($n = 21$) (with (c): $p = 0.0463$, $ES_{E1} = 0.756$)	41.6 ($n = 48$) (with (c): $p = 0.0153$, $ES_{E1} = 0.758$)	69.7 ($n = 20$)	1373
TF1RH (seconds)	79.9 ($n = 24$) (with (c): $p = 0.0000777$, $ES_{E1} = 1.21$)	63.3 ($n = 47$) (with (c): $p = 0.0000002$, $ES_{E1} = 1.36$)	214.6 ($n = 30$)	12433
TF1H (seconds)	127.7 ($n = 23$) (with (c): $p = 0.0373$, $ES_{E1} = 0.712$)	119.5 ($n = 36$) (with (c): $p = 0.00919$, $ES_{E1} = 0.771$)	226.9 ($n = 27$)	19423
ATBJ (seconds)	25.7 ($n = 25$) (with (c): $p = 0.0000024$, $ES_{E1} = 1.27$)	20.8 ($n = 50$) (with (c): $p \approx 0$, $ES_{E1} = 1.69$)	40.7 ($n = 50$)	139
NREJ (times)	8.04 ($n = 25$) (with (c): $p = 0.0153$, $ES_{E1} = 0.688$)	5.08 ($n = 50$) (with (c): $p = 0.0000317$, $ES_{E1} = 0.916$)	17.0 ($n = 50$)	169.5

assessment interface. Figure 1(a) and (b) show the Gold-relevant documents summed over the 25 PRI-Gold topics and 25 RND-Gold topics, respectively, while (c) and (d) show the Bronze-relevant documents summed over the 50 topics, for PRI-BronzeW and PRI-BronzeT, respectively. The graphs show document ranks in the pool files from 1 to 140, because 140 was the minimum pool size across the topic set as shown in Table 1: that is, below this rank, not every topic has a document to be judged. The following observations can be made.

- By comparing (a) and (b), it can be observed that Gold assessors tend to find more relevant documents near the top ranks of the PRI pool files, while no such tendency is observed for the RND files. This is consistent with the Bronze results of Sakai et al. [41, Figure 5]. Since the present study features Gold assessors and therefore the (highly) relevant documents identified by them can basically be considered correct (with some human errors—see Section 8), Figure 1(a) suggests that the PRI strategy is effective for putting likely relevant documents near the top ranks for the assessors. In other words, the pseudorelevant documents are in fact often truly relevant; note that it was not possible to make the same remark in Sakai et al. [41] as that study did not involve Gold assessors.
- The Bronze results of (c) and (d) are also consistent with Sakai et al. [41]: Bronze assessors also tend to find more (highly) relevant documents near the top ranks of the PRI pool files.

PRI's tendency to place likely relevant documents near the top of the document list probably explains why assessors tend to find the first (highly) relevant document more quickly with PRI pools than with RND pools.

4.2 RQ1.2: Gold vs. Bronze under a PRI Environment (Efficiency)

Table 4 addresses **RQ1.2** (*Which assessor type enables more efficient relevance assessments under a PRI environment?*) by comparing the efficiency statistics of the PRI-Gold, PRI-BronzeW, and PRI-BronzeT assessments, where the default sample size for the Bronze statistics is 50. We first note that the statistics of PRI-BronzeT are considerably larger than those of PRI-Gold and PRI-BronzeW. For example, it can be observed that the sample size of PRI-BronzeT for TJ1D is only 20, which means that we lost as many as 30 topics as a result of applying the 3-minute cap as described earlier. Despite this rigorous thresholding, the remarkable characteristics of the BronzeT assessors are clear: as indicated in the table, for every criterion, PRI-BronzeT is statistically significantly "less efficient" (or "more careful") than PRI-Gold and PRI-BronzeW (unpaired Tukey HSD test), whereas the difference between PRI-Gold and PRI-BronzeW is small and not statistically significant. In particular, in terms of ATBJ, our main efficiency criterion, the BronzeT assessors spent 40.7 seconds per document, while the BronzeW and Gold assessors spend only 20.8 seconds and 25.7 seconds per document, respectively.[14]

The high efficiency of the Gold assessors is not altogether surprising as they are the query owners and they *know* what they want. However, how can we explain the striking difference in characteristics between the BronzeW and BronzeT assessors? In Section 4.3, we will show that the assessors in each group (e.g., BronzeW) performed more or less similarly to one another in terms of efficiency. Moreover, in Section 5, we will show that none of the Bronze assessors are "outliers" in terms of inter-assessor agreement, which suggests that every assessor did a conscientious job. Nevertheless, recall that the BronzeW assessors are students, paid by the hour, whereas the BronzeT assessors are professional labellers. While the BronzeW assessors are let go after job completion, the BronzeT assessors continue to work for their company, and they have a reputation to upkeep. In short, we believe that the BronzeT assessors are more highly motivated, and probably have much more experience in labelling tasks.

Based on the above discussion, we conclude as follows. *Regarding the assessor efficiency under a PRI environment, what probably matters more than the assessor type (Gold vs. Bronze) is whether the Bronze assessors are highly motivated and/or experienced. If they are, they may spend substantially longer judgement times than Gold assessors do.*

4.3 Assessor-level Analysis (Efficiency)

Section 4.2 compared Gold assessors and Bronze assessors under the PRI environment in terms of efficiency, but the analysis assumes that the assessors within each assessor type share similar characteristics. However, all assessors are human, with varying traits, knowledge, and motivation. For example, while all seven authors of this paper (i.e., researchers involved in IR evaluation) served as Gold assessors, we imposed no particular control to ensure that this assessor group behaves homogeneously; by definition, "Gold" merely means that a topic creator is also the relevance assessor. This is why this section examines the assessor efficiency at the *assessor* level rather than the *assessor-type* level: how homogeneous are the assessors within each assessor type?

Figure 2 plots the individual efficiency statistics for Gold, BronzeT, and BronzeW assessors (after the aforementioned thresholding) under the PRI environment to enable assessor-level comparisons: Gold, BronzeT, BronzeW statistics are shown in gold, green, and red, respectively, with different symbols indicating different assessors. It can be observed that (a) the upper part of each graph is dominated by green (i.e., BronzeT statistics), and, more importantly, (b) the majority of the BronzeT assessors contribute to the said trend, as *different* green symbols in each graph visualise. For example, in Figure 2(d) for ATBJ, *every* BronzeT assessor has one or more data points above the 30-second horizontal grid. On the other hand, both the Gold and BronzeW statistics tend to lie beneath the BronzeT ones, although these two groups are not separable from each other. In other words, we do not see any "outlier" Gold assessor who behaves consistently differently compared to the others; we do not see any "outlier" BronzeW assessor either. This analysis suggests that our *assessor-type*-level analysis shown in Table 4 is of some value at least, since the members within each assessor type seem relatively homogeneous in the sense discussed as above, despite the fact that the Gold-Bronze distinction is solely based on whether the assessor is the topic creator or not.

[14] For reference, the Bronze assessors in the experiments of Sakai et al. [41, Table 2] spent 13.1-15.8 seconds per document ($n = 160$). However, these statistics are not directly comparable with ours, as not only the topics but also the target corpora are different: while our target corpus (for the WWW-4 topics) is Chuweb21, theirs (for the WWW-2 and WWW-3 topics) is ClueWeb12-B13.

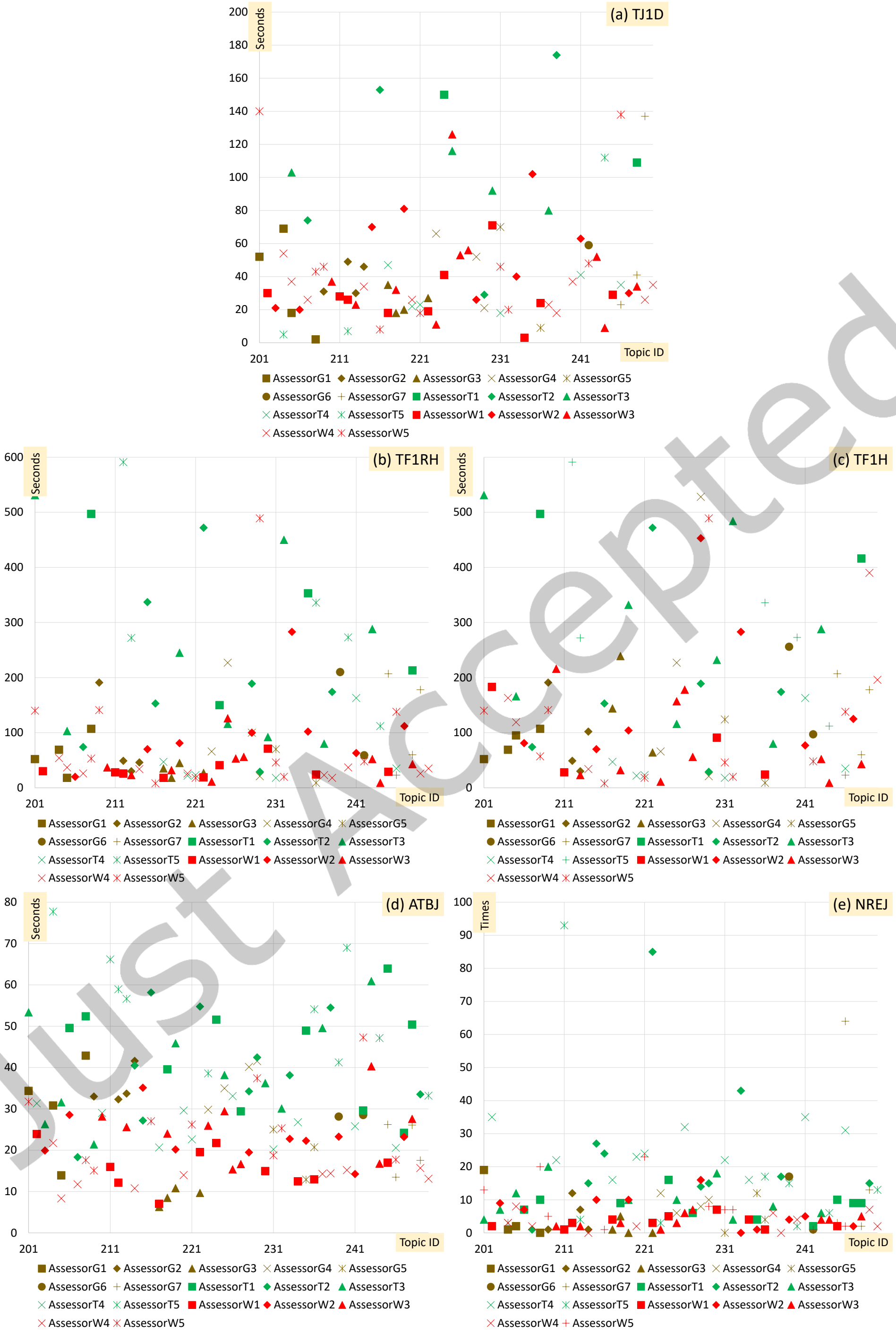

Fig. 2. Assessor-level comparison of efficiency statistics: Gold (in gold), BronzeT (in green), and BronzeW (in red) assessors.

Table 5. Inter-assessor agreement in terms of mean quadratic weighted Cohen's κ for each pair of assessment environments. None of the differences in means are statistically significant according to a Tukey HSD test for unpaired data. Residual variance for computing effect sizes: $V_{E1} = 0.03511$ [32].

Assessment environment pair	#topics	Mean κ
PRI-Gold vs. PRI-BronzeT	25	0.5324
PRI-BronzeW vs PRI-BronzeT	50	0.4575
RND-Gold vs. PRI-BronzeT	25	0.4568
PRI-Gold vs. PRI-BronzeW	25	0.4445
RND-Gold vs. PRI-BronzeW	25	0.4350

5 RQ2: INTER-ASSESSOR AGREEMENT

This section addresses **RQ2** (inter-assessor agreement) by examining the effect of assessor type (i.e., Gold vs. Bronze) and document ordering strategy (i.e., PRI vs. RND) on pairwise agreement in terms of quadratic weighted Cohen's κ [11, 34] for each topic, where the assessor labels are treated as 2 (highly relevant), 1 (relevant), and 0 (nonrelevant).

Table 5 shows the inter-assessor agreement for each pair of assessment environments (i.e., combinations of document ordering strategy and assessor type). Recall that PRI and RND cannot be compared under a Bronze environment in our study; that was already covered in Sakai et al. [40, 41]. While the mean κ's vary somewhat, none of the differences in mean κ are statistically significant according to a Tukey HSD test for unpaired data at the 5% significance level. Nevertheless, the two Gold-BronzeT mean agreements are higher than the two Gold-BronzeW mean agreements, which is in line with our assessor efficiency results that showed that the BronzeT assessors tend to be "more careful" than the BronzeW assessors. For example, the effect size between "PRI-Gold vs. PRI-BronzeT" and "PRI-Gold vs. PRI-BronzeW" is $(0.5324 - 0.4445)/\sqrt{0.03511} = 0.469$. These results, despite the lack of statistical significance, suggest that whether the Bronze assessors are highly motivated and/or experienced may potentially have a nonnegligible effect on the Gold-Bronze agreements. Assuming that the Gold assessments are correct, it is possible that the BronzeT assessments may be of higher quality than the BronzeW ones.

Tables 6 and 7 examine the per-topic κ's at the individual assessor level, in order to demonstrate that there are no clear "outlier" assessors who behave very differently from others to heavily affect the overall inter-assessor agreements. For example, the "AssessorG1" row of Table 6(a) compares the labels of AssessorG1 with those from Assessor{W5,W1,W2,W4,W4,W2,W4,W5} to compute the mean τ over topics 0201-0208 (See Table 1 "Bronze" column); similarly the "AssessorW1" row of Table 7(a) compares the labels of AssessorW1 with those from Assessor{G1,G2,G2,G3,G3,G4,G5,G5,G5,G7} to compute the mean τ over topics 0202, 0211, 0212, 0217, 0222, 0224, 0230, 0234, 0236, 0245 (See Table 1). It can be observed that the mean τ's for different assessors within each group (i.e., Gold, BronzeW, or BronzeT) do not vary drastically, which suggests the within-group homogeneity of assessment *reliability*. Thus, along with the assessor-level *efficiency* results shown in Figure 2, these relatively similar assessor-level agreements suggest that our *assessor-type* level analyses in the present study are useful to some extent.

In addition, by comparing the two columns of Table 6 as well as the top and bottom sections of Table 7(a), it can be observed that the Gold-BronzeT agreements tend to be higher than the Gold-BronzeW agreements even at the assessor level. In particular, in Table 6, for every assessor except for AssessorG4, the mean agreement with BronzeT is higher than that with BronzeW on average. Again, assuming that the Gold assessments are correct, these assessor-level results also suggest that the BronzeT assessments may be more accurate than the BronzeW ones.

Table 6. Mean per-topic inter-assessor agreement for each gold assessor in terms of quadratic weighted Cohen's κ ($n = 8$ topics for Gold01; $n = 7$ topics for the others). For example, the labels of AssessorG1 are compared with those given by the BronzeW and BronzeT assessors.

Assessor	(a) Mean κ (with BronzeW)	(b) Mean κ (with BronzeT)
AssessorG1	0.393	0.473
AssessorG2	0.438	0.583
AssessorG3	0.329	0.366
AssessorG4	0.504	0.473
AssessorG5	0.453	0.486
AssessorG6	0.414	0.463
AssessorG7	0.554	0.623

Table 7. Mean per-topic inter-assessor agreement for each bronze assessor in terms of quadratic weighted Cohen's κ ($n = 10$ topics). For example, the labels of AssessorW1 are compared with those given by the Gold and BronzeT assessors.

Assessor	(a) Mean κ (with Gold)	(b) Mean κ (with BronzeT)
AssessorW1	0.395	0.450
AssessorW2	0.460	0.459
AssessorW3	0.490	0.444
AssessorW4	0.426	0.428
AssessorW5	0.428	0.507
Assessor	(a) Mean κ (with Gold)	(b) Mean κ (with BronzeW)
AssessorT1	0.549	0.476
AssessorT2	0.480	0.485
AssessorT3	0.462	0.395
AssessorT4	0.481	0.500
AssessorT5	0.501	0.432

In addition to the topic-level inter-assessor agreements discussed above in terms of κ statistics, we also investigated *document*-level inter-assessor agreements in terms of raw document counts. Table 8 shows a breakdown of all Gold-Bronze per-document disagreements by whether the Gold assessor says (highly) relevant or not, and by whether the Gold assessor used a PRI file or a RND file. For example, Part (I)(a) shows that, under the PRI environment, there are 10 "serious disagreements" where the Gold assessor says highly relevant (L2) while both of the Bronze assessors say nonrelevant (L0); the total number of disagreements is 383 (out of 1,592 documents from the PRI-Gold topics, where the Gold assessor gave either an L2 or an L1). In contrast, Part (I)(b) shows that, for topics where the Gold assessor used a RND file (while the Bronze assessor used a PRI file as always), the number of disagreements is noticeably higher, with a total of 613 (out of 1,587 documents from the RND-Gold topics, where the Gold assessor gave either an L2 or an L1). Similarly, Part (II) shows the disagreements where the Gold assessor says nonrelevant: these cases can be regarded as noise introduced by the Bronze assessors, assuming that the Gold assessors are always correct.

It can be observed that Table 8 has more disagreements on the right than on the left. In particular, the difference in Section (I) is statistically significant according to an equal proportion test (383/1592 vs. 613/1587, $p < 2.2\text{e-}16$), although that in Section (II) is not. This probably reflects the impact of the document ordering strategy: as we have seen in Figure 1, under the PRI environment, the top ranked documents tend to be rated as relevant regardless of the assessor type, and therefore we obtain relatively fewer Gold-Bronze disagreements in the left part of the table.

Table 8. Gold-Bronze disagreement statistics: the numbers shown are topic-document pairs.

(I) Gold says (highly) relevant					
(a) Docs from PRI-Gold topics (1,592)			(b) Docs from RND-Gold topics (1,587)		
Gold label	Bronze labels	#topicdocs	Gold label	Bronze labels	#topicdocs
L2	both L0	10	L2	both L0	22
L2	one L0, one L1/L2	78	L2	one L0, one L1/L2	106
L1	both L0	76	L1	both L0	127
L1	one L0, one L1/L2	219	L1	one L0, one L1/L2	358
Total		383	Total		613
(II) Gold says nonrelevant					
(a) Docs from PRI-Gold topics (3,537)			(b) Docs from RND-Gold topics (3,617)		
Gold label	Bronze labels	#topicdocs	Gold label	Bronze labels	#topicdocs
L0	both L2	**51**	L0	both L2	50
L0	one L2, one L1	**167**	L0	one L2, one L1	184
L0	both L1	221	L0	both L1	222
Total		439	Total		456

That is, while the differences in the *topic-level* Gold-Bronze agreements in terms of mean κ's are not statistically significant (Table 5), we have evidence that *document-level* Gold-Bronze agreements are affected by the document ordering strategy.[15]

Section (I) of Table 8 quantifies the fact that Bronze assessors can miss truly relevant documents (as defined by the Gold assessors), which is not altogether surprising. What we find more worrying is the "noise" introduced by the Bronze assessors, represented by Section (II) of the same table. We have seen in Table 2 that Bronze assessors tend to find more relevant documents than Gold assessors, and this was why we set up **RQ5** (liberal Bronze assessors). The document counts shown in Table 8 Section (II) are the exact cause of the liberal nature of the Bronze assessments. In particular, Table 8(II)(a) shows that, even when both the Gold and the Bronze assessors are in the PRI environment, there were a total of $(51 + 167) = 218$ serious disagreements where the two Bronze assessors said either (L2, L2) or (L2, L1) even though the Gold assessor said L0. In Section 8, we shall, as Gold assessors, closely re-examine these 218 documents to address **RQ5**: why do Bronze assessors tend to be liberal?

Based on the above discussions, our answer to **RQ2** is as follows. *We obtain fewer Gold-Bronze disagreements in terms of document counts when both Gold and Bronze assessors are in the PRI environment than when only the Gold assessors are in the RND environment. In this sense, the document ordering strategy affects inter-assessor agreement.* Also, while the topic-level inter-assessor agreement results in terms of mean κ's are not statistically significant, the effect sizes from Table 5 suggest that BronzeT assessments may be of higher quality than the BronzeW ones.

6 RQ3: SYSTEM RANKING SIMILARITY

This section addresses **RQ3** (system ranking similarity) by ranking the 18 NTCIR-16 WWW-4 runs [38, 39] by mean effectiveness measure scores according to Gold, BronzeW and BronzeT versions of the qrels files. While we have 25 PRI-Gold and 25 RND-Gold topics as we have discussed earlier, hereafter we exclude one RND-Gold topic (Topic 0203) from our experiments, as this topic does not have any relevant documents in the BronzeW qrels file.

[15] Note that the mean κ treats each *topic* equally; if there are many documents to be judged for a topic, the contribution of each document to the κ for that topic becomes small.

Table 9. Kendall's τ (with 95%CIs, $n = 18$) between a system ranking based on means over the 25 PRI-Gold topics and one based on means over the 24 RND-Gold topics, where both use the same qrels file (Gold, BronzeW, or BronzeT).

Measure	(a) Gold (PRI vs. RND)	(b) BronzeW (PRI vs. PRI)	(c) BronzeT (PRI vs. PRI)
nDCG	0.621 [0.363, 0.791]	0.503 [0.204, 0.716]	0.686 [0.457, 0.830]
Q	0.608 [0.345, 0.783]	0.490 [0.188, 0.708]	0.686 [0.457, 0.830]
nERR	0.490 [0.188, 0.708]	0.327 [−0.007, 0.595]	0.660 [0.419, 0.814]
iRBU	0.542 [0.255, 0.741]	0.294 [−0.043, 0.571]	0.634 [0.381, 0.798]

That is, when we average over the entire topic set, we use 49 topics; when we average over the RND-Gold topic set, we use 24 topics.[16]

6.1 RQ3.1: PRI vs. RND under a Gold Environment (System Ranking Similarity)

Table 9(a) addresses **RQ3.1** (*How similar are PRI-based and RND-based system rankings under a Gold environment?*) by comparing two system rankings using the Gold qrels file in terms of Kendall's τ with 95%CIs [22, 26]: the first ranking is based on mean effectiveness scores over the 25 PRI-Gold topics, while the second ranking is based on mean effectiveness scores over the 24 RND-Gold topics, for each of the four official evaluation measures. It can be observed that the Gold rank correlations are reasonable despite the use of two different topic sets, each judged under a different environment, i.e., PRI or RND. The τ's are lower compared to the Bronze rank correlations reported in Sakai et al. [41, Table 7][17], but it should be noted that their PRI-RND system ranking comparisons were based on a *common* topic set. Hence, our answer to **RQ3.1** is: *PRI-based and RND-based system rankings (with two disjoint topic sets) under a Gold environment are reasonably similar.* Taken together with the Bronze-based results of Sakai et al. [41], *the document ordering strategy (regardless of assessor type) do affect system rankings substantially, but not drastically.*

For completeness, Table 9(b) and (c) show the τ's with each of the two Bronze qrels files, where one ranking is based on the PRI-Gold topics and the other is based on the RND-Gold topics; but note that the Bronze assessors were in a PRI environment even for the latter topic set.[18] It can be observed that while the BronzeW τ correlations with nERR and with iRBU are not even statistically significantly correlated, the BronzeT system rankings are much more stable across the two document ordering strategies, for each evaluation measure. These rank correlation results also suggest that the BronzeT assessments are more reliable than BronzeW assessments.

6.2 RQ3.2: Gold vs. Bronze under a PRI Environment (System Ranking Similarity)

Table 10 addresses **RQ3.2** (*How similar are Gold-based and Bronze-based system rankings under a PRI environment?*) by comparing the Gold, BronzeW, and BronzeT system rankings, where all rankings are based on the same 25 PRI-Gold topics. It can be observed that the Gold-BronzeT correlations are particularly high, with Q-measure achieving a τ of 0.804 and nDCG achieving 0.739. In this sense, the BronzeT qrels file is a reasonable substitute for the Gold qrels file. This is probably because the BronzeT assessments are relatively accurate, as our inter-assessor agreement results suggested (Table 5). In contrast, the Gold-BronzeW correlations and the BronzeW-BronzeT correlations are very similar, with nERR and iRBU failing to achieve statistically significant correlations in both

[16] The official "Bronze-All" results of the NTCIR-16 WWW-4 task used the entire 50 topics: the Bronze-All qrels file was constructed by combining the BronzeW and BronzeT relevance assessments so every topic had some relevant documents [38, 39].

[17] The PRI-RND correlations in Sakai et al. [41] (using a common topic set in a Bronze environment, $n = 36$) were in the range of 0.784-0.908 for nDCG, 0.746-0.922 for Q, 0.733-0.857 for nERR, and 0.762-0.852 for iRBU.

[18] Comparing these with (a) is not very useful since (a) is the only setting where a PRI-based and RND-based rankings are compared.

Table 10. Kendall's τ (with 95%CIs, $n = 18$) between two system rankings based on different qrels files (Gold, BronzeW, or BronzeT), where both use the same 25 PRI-Gold topics.

Measure	(a) Gold vs. BronzeW	(b) Gold vs. BronzeT	(c) BronzeW vs. BronzeT
nDCG	0.490 [0.188, 0.708]	0.739 [0.538, 0.860]	0.516 [0.221, 0.725]
Q	0.438 [0.123, 0.673]	0.804 [0.643, 0.897]	0.477 [0.171, 0.699]
nERR	0.314 [−0.021, 0.586]	0.516 [0.221, 0.725]	0.327 [−0.007, 0.595]
iRBU	0.248 [−0.093, 0.537]	0.549 [0.264, 0.746]	0.307 [−0.029, 0.581]

cases. Hence, assessor type does not appear to be the most important factor: probably a more important question is "*Which* Bronze qrels file?"

In summary, to answer **RQ3.2**: *under the PRI environment, the Gold-Bronze rank correlations (with the same topic set) can be high, but this depends substantially on the quality of the Bronze assessments.*

Finally, both Tables 9 (different topic sets, same qrels file) and Table 10 (different qrels files, same topic set) suggest that nDCG and Q provide more stable system rankings than nERR and iRBU do. This is probably due to the use of the *diminishing-return* decay function used in nERR and iRBU, which makes the measures "shallow" especially when there are highly relevant documents near the top of the ranked list [36].

7 RQ4: SYSTEM RANKING CONSISTENCY (ROBUSTNESS TO THE CHOICE OF TEST DATA)

This section addresses **RQ4** (system ranking consistency) using the procedure described in Sakai [36]: for each qrels file, its robustness to the choice of test topics is examined by randomly splitting the topic set in half, creating two system rankings based on the two subsets, and computing the rank correlation.[19] The random splitting is conducted $B = 1,000$ times so that a mean Kendall's τ score is obtained for that qrels file; we want the mean τ to be high because we do not want the system ranking to change drastically just because the test sample (i.e., topic set) has been replaced by another. Note that this *system ranking consistency* evaluation was not considered in Sakai et al. [41].

7.1 RQ4.1: PRI vs. RND under a Gold Environment (System Ranking Consistency)

Table 11 addresses **RQ4.1** (*How robust to the choice of test data are PRI-based and RND-based system rankings under a Gold environment?*) by comparing system ranking consistencies for Gold assessments: the PRI-based and RND-based results are obtained by splitting the PRI-Gold topics and the RND-Gold topics in half, respectively. For example, "PRI-Gold-nDCG" represents the experiment where, in each trial, the 25 PRI-Gold topics are split into subsets of size 13 and 12, and two nDCG-based system rankings are obtained with these subsets. Statistically significant differences at the 5% significance level (with a paired randomised Tukey HSD test) are also indicated. For example, RND-Gold-nDCG statistically significantly outperforms all others and therefore is the most reliable combination of document ordering strategy and evaluation measure. More generally, for each evaluation measure (with the exception of iRBU), the RND-based experiment statistically significantly outperforms the corresponding PRI-based experiment. Because the PRI-Gold topic set and the RND-Gold topic set are disjoint and different to begin with, it is not entirely clear whether the superiority of the RND-Gold topic set results is due to the document ordering strategy or to different levels of variations in each topic set. For example, if the topic set contains topics that each rank the systems very differently, then this will inevitably lower the system ranking consistency across two topic subsets. Hence, regarding **RQ4.1**, we can only say the following. *In our experiments, RND-Gold system rankings were generally more robust to the choice of test data than the PRI-Gold system rankings. However, it is not*

[19] This method is related to the *swap method* [52], but quantifies the discrepancy between topic sets in terms of Kendall's τ between two entire system rankings rather than the "swap rate."

Table 11. System ranking consistency in terms of mean τ for Gold assessments. The PRI-based and RND-based results are obtained by splitting the PRI-Gold topics and the RND-Gold topics in half, respectively. Statistical significance is determined based on a randomised Tukey HSD test [32] at the 5% significance level with $B = 1,000$ trials. The residual variance for computing effect sizes is $V_{E2} = 0.0175$ [32]. Upon one reviewer's request, the sample standard deviation of τ for each experiment is also shown.

pool type/assessor type/measure	mean τ	statistically significantly outperforms ($\alpha = 0.05$)	Sample standard deviation
a. RND-Gold-nDCG	0.5414	b,c,d,e,f,g,h	0.1286
b. RND-Gold-Q	0.4980	c,d,e,f,g,h	0.0907
c. RND-Gold-nERR	0.4548	d,e,f,g,h	0.1299
d. PRI-Gold-nDCG	0.4165	e,f,g,h	0.1264
e. PRI-Gold-Q	0.3726	f,g,h	0.1391
f. RND-Gold-iRBU	0.2946		0.1676
g. PRI-Gold-iRBU	0.2909		0.1817
h. PRI-Gold-nERR	0.2878		0.1673

clear whether the difference is due to the document ordering strategy or to different levels of variations in each topic set. Experiments with larger topic sets are needed in order to reduce the likelihood of the latter situation. Note that it is not possible to compare PRI-Gold and RND-Gold environments using a *common* topic set, unless each Gold assessor is somehow made to process both pool types for the same topics.

7.2 RQ4.2: Gold vs. Bronze under a PRI Environment (System Ranking Consistency)

Table 12 addresses **RQ4.2** (*How robust to the choice of test data are Gold-based and Bronze-based system rankings under a PRI environment?*) by comparing the system ranking consistencies of the Gold, BronzeW, BronzeT system rankings using the 25 PRI-Gold topics. Note that the PRI-Gold results are duplicated from Table 11. The following observations can be made.

- The two top performers (PRI-BronzeT-{Q, nDCG}) statistically significantly outperform all others, *including* PRI-Gold-{Q, nDCG}. That is, for these two evaluation measures, BronzeT is actually superior to Gold in terms of system ranking consistency.
- The four BronzeW results are the worst performers. In particular, when coupled with the BronzeW assessments, the nERR and iRBU system rankings break down completely as a result of using completely different topic sets. Thus, BronzeW underperforms BronzeT from the viewpoint of system ranking consistency as well.

From the above results, our answer to **RQ4.2** is as follows. *The primary factor that affects system ranking consistency under the PRI environment is how reliable the Bronze assessments are rather than the assessor type; a high-quality Bronze qrels file can be more robust to the choice of test topics than a Gold qrels file.* Whether it is possible to close this gap by introducing some quality control over Gold assessors is a question left for future work.

As a final remark on Table 12, the complete lack of robustness to the choice of test data for nERR and iRBU based on the BronzeW assessments is consistent with the system ranking similarity results discussed in Section 7 (Tables 9 and 10).

Table 12. System ranking consistency in terms of mean τ under the PRI environment. The results are obtained by splitting the 25 PRI-Gold topics in half. Statistical significance is determined based on a randomised Tukey HSD test [32] at the 5% significance level with $B = 1,000$ trials. The residual variance for computing effect sizes is $V_{E2} = 0.0185$ [32]. The sample standard deviation of τ for each experiment is also shown.

Pool type/assessor type/measure	Mean τ	Statistically significantly outperforms ($\alpha = 0.05$)	Sample standard deviation
a. PRI-BronzeT-Q	0.5121	c,d,e,f,g,h,i,j,k,l	0.1490
b. PRI-BronzeT-nDCG	0.4934	c,d,e,f,g,h,i,j,k,l	0.1385
c. PRI-Gold-nDCG	0.4165	d,e,f,g,h,i,j,k,l	0.1264
d. PRI-Gold-Q	0.3726	f,g,h,i,j,k,l	0.1391
e. PRI-BronzeT-iRBU	0.3473	f,g,h,i,j,k,l	0.1620
f. PRI-Gold-iRBU	0.2909	k,l	0.1817
g. PRI-Gold-nERR	0.2878	k,l	0.1673
h. PRI-BronzeT-nERR	0.2802	k,l	0.1647
i. PRI-BronzeW-nDCG	0.2697	k,l	0.1721
j. PRI-BronzeW-Q	0.2571	k,l	0.1616
k. PRI-BronzeW-iRBU	−0.1084		0.1429
l. PRI-BronzeW-nERR	−0.1131		0.1571

8 RQ5: LIBERAL BRONZE ASSESSORS

This section addresses **RQ5** (liberal bronze assessors) through a small-scale additional relevance assessment experiment where the Gold assessors (i.e., authors of this paper) re-assessed the 218 documents shown in bold in Table 8: these are the documents originally labelled as L0 by the Gold assessors but were labelled as either (L2, L2) or (L2, L1) by the two Bronze assessors, all under the PRI environment. Our objective was to find out why Bronze assessors find *additional* relevant documents: are they truly relevant documents that even the Gold assessors originally missed, or are they just noise, due to, say, lack of understanding of the topic? By definition, only the Gold assessors can answer these questions; hence our additional effort.

Table 13 shows the *content* ("title" in TREC parlance) and *description* fields of the 23 topics that are involved in this rejudging experiment. On the relevance assessment interface, only the documents that needed to be re-examined were presented to the Gold assessors, in the original PRI-based order. Table 14 shows the distribution of new Gold labels for each assessor, and Table 15 breaks down the table further by topic. The following can be observed from these tables.

- Of all the documents originally judged nonrelevant by the Gold assessor but judged (highly) relevant by both of the Bronze assessors, almost one half (31.2% + 15.6% = 46.8%) are truly relevant or highly relevant according to the re-examination. That is, the Gold assessors did actually miss some relevant documents in the original round.
- For as many as 20 topics out of the 23, at least one document was newly recognised as (highly) relevant by the Gold assessor. (The three exceptions are 0204, 0205, and 0239.)

The Gold assessors took down notes when they re-assessed their documents. We examined their remarks on the newly recognised highly relevant documents, and many of them were to the effect of "I don't know why I missed this" or "It should have been labelled highly relevant in the first round." That is, the Gold assessors realised they were wrong. In addition, one assessor made remarks such as "it is about chicken recipes, but is difficult to say whether they are chicken *breast* recipes." ("0245 chicken breast recipes") and "the main idea of the document is not about the price, but it does mention the price" ("0248 PS5 price"). Thus, there were cases where the Gold assessor remarked that Bronze assessors "may be right."

The above discovery of these new Gold-relevant documents is not entirely surprising, and does not necessarily imply that the Gold assessors were quite careless. We already know that how the pooled documents are presented can affect relevance assessments (see Figure 1); recall that in this additional experiment, only the 218 serious disagreements (218/23 = 9.5 documents per topic on average; see Table 15) were rejudged by the Gold assessors; this is very different from the original situation where they had to process 206.7 documents per topic on average (See Section 3.3), even though both used the PRI ordering scheme. That is, given these substantially smaller lists of documents to judge, it is only natural that each Gold assessor paid closer attention to each document. Furthermore, there is one more potential factor that might have enhanced the Gold assessors' attention: while the Gold assessors were given no prior information about the pooled documents in the original assessment phase, they *knew* that the documents in the additional assessment phase were those judged relevant by Bronze assessors.[20] Hence the Gold assessors may have been curious and motivated than before, thinking: "*Why on earth did they think these documents are relevant?*"

In summary, our answer to **RQ5** is as follows. *Bronze assessors tend to be liberal not only because they label some nonrelevant documents as relevant, but also because they find some relevant documents that even the Gold assessors miss. In our experiments, of the documents judged nonrelevant by the Gold assessor contrary to the two Bronze assessors, almost one half were truly relevant according to the Gold assessors' own reconsiderations.*

For the reasons discussed earlier, it is not entirely surprising that Gold assessors can find additional relevant documents in a second-round assessment phase. Moreover, it should be stressed that "Gold" merely means that the topic creator is also the assessor; it does *not* mean that that person achieves 100% precision and 100% recall; just like Bronze assessors, they are human. While introducing some quality control over Gold assessors may improve on the above "one-half" situation, we highly doubt that the additional discovery of relevant documents can be completely eliminated.

[20] Because the present authors are the Gold assessors, it was not possible to conduct an additional experiment where the Gold assessors are not aware of its purpose.

Table 13. Content and description fields of the 23 topics that were involved in the rejudging experiment.

0201	Timnit Gebru Google	I want to know the details regarding Google's firing of Dr. Timnit Gebru.
0204	Tokyo olympics coronavirus athlete	Which athletes or former athletes expressed concerns about holding the Tokyo Olympics amid COVID-19?
0205	thomas dolby songs	I'm collecting information about songs released by Thomas Dolby.
0208	cultural appropriation cases	I want to read about specific situations considered by some to be cultural appropriation. Who have been accused?
0209	YOLO	what is the meaning of "YOLO"?
0212	chiffon cake recipe	I want to find some recipes for making a chiffon cake
0213	dark chocolate health	whether eating dark chocolate will benefit our health or not?
0214	cat lily poisonous	Is lily poisonous to cats?
0217	inventor of the Web	Who is the inventor of the World Wide Web?
0218	deep Web	What is the deep Web?
0219	What is SEO?	What is Search Engine Optimization?
0222	NoSQL	I'm looking for a defintion of NoSQL and what is it
0223	czechoslovakia divide reason 1993	You want to know why Czechoslovakia was divided into two countries
0225	signifier saussure theory	You want to know the meaning of term "signifier" in linguist Saussure's theory
0228	block chain crypto	You want to know the relationship between block chain and crypto currencies
0229	side effect pfizer	You are going to have the jab and you want to know the possible side effect of pfizer
0231	beautiful mess lyrics	You want to find the lyrics of the song "beautiful mess"
0236	tennis score rules	You want to know the scoring rules of tennis matches
0239	malawi 2019 presidential elections	You heard that in 2019 Malawi had presidential elections that were invalidated due to fraud in votes counting. You want to know more about this.
0242	kiruna moving town	You heard that Kiruna, a small town in Sweden, needs to be relocated. You would like to have more information about this town.
0245	chicken breast recipes	You want to learn some chicken breast recipes
0246	best game nintendo switch	You want to know some popluar [*sic*] game titles on Nintendo Switch
0248	PS5 price	You want to know the price of a PlayStation 5

Table 14. Gold assessors' new labels after rejudging the 218 documents.

Assessor	#rejudged	L0	L1	L2
AssessorG1	15	86.7% (13)	13.3% (2)	0.0% (0)
AssessorG2	34	20.6% (7)	67.6% (23)	11.8% (4)
AssessorG3	93	69.9% (65)	17.2% (16)	12.9% (12)
AssessorG4	26	34.6% (9)	46.2% (12)	19.2% (5)
AssessorG5	4	0.0% (0)	25.0% (1)	75.0% (3)
AssessorG6	15	80.0% (12)	13.3% (2)	6.7% (1)
AssessorG7	31	32.3% (10)	38.7% (12)	29.0% (9)
All	218	53.2% (116)	31.2% (68)	15.6% (34)

Table 15. Gold assessors' new labels after rejudging the 218 documents: breakdown by topic.

Assessor	Topic ID	#rejudged	L0	L1	L2
AssessorG1	0201	1	0	1	0
	0204	3	3	0	0
	0205	1	1	0	0
	0208	10	9	1	0
AssessorG2	0209	2	0	2	0
	0212	6	1	3	2
	0213	23	5	16	2
	0214	3	1	2	0
AssessorG3	0217	2	1	1	0
	0218	42	32	5	5
	0219	39	27	7	5
	0222	10	5	3	2
AssessorG4	0223	2	1	1	0
	0225	2	1	1	0
	0228	6	2	3	1
	0229	16	5	7	4
AssessorG5	0231	3	0	1	2
	0236	1	0	0	1
AssessorG6	0239	6	6	0	0
	0242	9	6	2	1
AssessorG7	0245	2	0	0	2
	0246	14	9	5	0
	0248	15	1	7	7
All		218	116	68	34

9 CONCLUSIONS

We examined two factors that may affect the relevance assessments of depth-k pools, which in turn may affect the relative evaluation of different IR systems. The first factor is the document ordering strategy for the assessors, namely, PRI and RND, following the work of Sakai et al. [41] where only Bronze assessors were involved in the experiments. The second factor is assessor type, i.e., Gold or Bronze. Studying the effect of assessor type on relevance assessments is important because (a) in most offline evaluation of IR systems, it is assumed that Bronze assessments are good substitutes for Gold assessments, even though it is usually the latter we actually want; and (b) several studies have suggested that the above assumption often does not hold (See Section 2.1). We believe that our study is unique in that the authors of this paper are the Gold assessors, which enabled us to closely examine why Bronze assessments differ from the Gold ones.

Our answers to the six main research questions are summarised below.

RQ1 is about assessor efficiency. **RQ1.1** compared PRI and RND with Gold assessments, and our conclusion is: *under the Gold environment, there is no substantial difference between PRI and RND in terms of time spent for judging each document, although assessors tend to find the first (highly) relevant document more quickly with PRI files.* These results for Gold assessors are generally in line with those for Bronze assessors [41, Table 2]. In addition, our close examination of the Gold assessments in the PRI environment suggests that the pseudorelevant documents suggested by the PRI approach are often indeed relevant. **RQ1.2** compared Gold and Bronze assessors under the PRI environment, and our conclusion is: *what probably matters much more than the assessor type (Gold vs. Bronze) is how the Bronze assessors are trained and/or well-motivated. If they are, they may spend substantially longer judgement times than Gold assessors do.*

RQ2 is about the possible impact of the document ordering strategy (PRI or RND) on inter-assessor agreement. Our conclusion is: *We obtain fewer Gold-Bronze disagreements in terms of document counts when both Gold and Bronze assessors are in the PRI environment than when only the Gold assessors are in the RND environment. In this sense, the document ordering strategy affects inter-assessor agreement.* Also, while the differences in the topic-level inter assessor agreements in terms of mean κ's were not statistically significant in our experiments, the effect sizes suggest that BronzeT assessments may resemble the Gold assessments more than the BronzeW assessments do.

RQ3 is about system ranking similarity. **RQ3.1** compared PRI-based and RND-based system rankings with Gold assessments, and our conclusion is: *PRI-based and RND-based system rankings (with two disjoint topic sets) under a Gold environment are reasonably similar.* Taken together with the Bronze-based results of Sakai et al. [41], *the document ordering strategy (regardless of assessor type) do affect system rankings substantially, but not drastically.* **RQ3.2** compared Gold and Bronze system rankings under the PRI environment, and our conclusion is: *the Gold-Bronze rank correlations (with the same topic set) can be high under the PRI environment, but this depends substantially on the quality of the Bronze assessments.*

RQ4 is about system ranking consistency, or the robustness of the system ranking to the choice of test topics. **RQ4.1** compared the system ranking consistency of PRI-based and RND-based system rankings with Gold assessments, and our conclusion is: *in our experiments, RND-Gold system rankings were generally more robust to the choice of test data than the PRI-Gold system rankings. However, it is not clear whether the difference is due to the document ordering strategy or to different levels of variations in each topic set, i.e., whether and how each topic in the topic set rank systems similarly or very differently.* Experiments with larger topic sets are needed in order to reduce the likelihood of the latter situation. **RQ4.2** compared Gold-based and Bronze-based system rankings under a PRI environment, and our conclusion is: *the primary factor that affects system ranking consistency under the PRI environment is how reliable the Bronze assessments are rather than the assessor type; a high-quality Bronze qrels file can be more robust to the choice of test topics than a Gold qrels file.* Whether it is possible to close this gap by introducing some quality control over Gold assessors is a question left for future work.

Our results for **RQ3** and **RQ4** also showed that system rankings based on nERR and iRBU are more fragile than those based on nDCG and Q. This is probably because both nERR and iRBU rely on a decay function which generally makes the measures "shallow" [36].

Finally, **RQ5** investigated why Bronze assessors tend to find more relevant documents compared to the Gold assessors. Our conclusion is: *Bronze assessors tend to be liberal not only because they label some nonrelevant documents as relevant, but also because they find some relevant documents that even the Gold assessors miss. In our experiments, of the documents judged nonrelevant by the Gold assessor contrary to the two Bronze assessors, almost one half were truly relevant according to the Gold assessors' own reconsiderations.* As we have discussed in Section 8, a second-round assessment environment is inherently different from the original assessment environment (e.g., in terms of the number of documents to assess and the prior knowledge about the documents), and therefore it is not entirely surprising that the Gold assessors find some additional relevant documents. Recall that we made no special attempt at controlling the quality of the Gold assessors:[21] while introducing some quality control may improve on the above "one-half" situation, we highly doubt that the additional discovery of relevant documents can be completely eliminated.

A high-level summary of the above results is that highly motivated/experienced Bronze assessors may be as reliable as Gold assessors in several ways, and that Gold-Bronze disagreements do not necessarily mean that the Bronze assessors are wrong. Hence, budget permitting, it is probably beneficial to hire highly-motivated Bronze assessors even when Gold assessors are available. Such an approach would make the test collection less incomplete, and therefore should enhance its reusability at least to some extent. Both Gold and Bronze assessors are human: neither are perfect, but they can complement each other.

In our future work, we would like to closely examine the relationship between the assessment quality and the *levels* of Bronze assessors' motivation and experience, as the present study only had BronzeW (students, part-time) and BronzeT (professional labellers working for a company). In addition, introducing some control over Gold assessors for quality assurance probably deserves some investigation. Furthermore, we would like to extend our investigation of document ordering and assessor type effects for constructing new types of test collections, such as those for a web search task that considers group fairness.[22]

ACKNOWLEDGEMENTS

We thank Yiqun Liu (Tsinghua University, P.R. China), Ian Soboroff (NIST, USA), and Junjie Wang (Waseda University) for working with us as the NTCIR-16 WWW-4 task organisers [38]. We also thank the participants of the WWW-4 task [48, 49, 56].

This paper was partially supported by the the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 893667.

APPENDIX – RQ6: ROBUSTNESS TO NEW SYSTEMS

Sakai et al. [41] compared their Bronze-based qrels in terms of *robustness to new systems* using the *Leave One Team Out* (LOTO) method [31]. That is, from the original qrels file, *unique contributions* from one particular participating team T (which in general submits multiple runs) are *left out* from the original qrels file to form a lo-T qrels file, and the system rankings before and after the removal are compared in terms of Kendall's τ. If the ranking according to lo-T qrels file is similar to the one according to the original qrels file, this means that the runs submitted by T can be evaluated fairly even though the unique contributions from T are being treated as nonrelevant; on the other hand, if the two rankings differ substantially, this is usually because the runs from T

[21] An implicit motivation that the Gold assessors had in common was that they, as the authors of this paper, all wanted this work to get published!

[22] http://sakailab.com/fairweb1/

Table 16. Unique contributions of each team and the size of the corresponding LOTO qrels file. The numbers in parentheses show the subtotals over the 25 PRI-Gold and 25 RND-Gold topics.

Team left out	#Runs	Unique contributions	(PRI/RND)	#Topic-document pairs in LOTO qrels	(PRI/RND)
ORG	2	24	(10/14)	10,309	(5,119/5,190)
KASYS	6	185	(93/92)	10,148	(5,036/5,112)
SLWWW	5	2932	(1,467/1,465)	7,401	(3,662/3,739)
THUIR	5	1793	(854/939)	8,540	(4,275/4,265)

Table 17. Robustness in terms of Kendall's τ (with 95%CIs, $n = 18$) when compared to the system ranking before leaving out a team. The PRI-Gold results are based on the mean scores over the 25 PRI-Gold topics; the RND-Gold results are based on the mean scores over the 24 RND-Gold topics. In each column (for each evaluation measure), the higher value is indicated in **bold**.

	lo-ORG	lo-KASYS	lo-SLWWW	lo-THUIR
PRI-Gold-nDCG	1	1	0.647 [0.400, 0.806]	**1**
RND-Gold-nDCG	1	1	**0.869** [0.754, 0.932]	0.961 [0.925, 0.981]
PRI-Gold-Q	1	1	0.569 [0.291, 0.758]	**0.922** [0.850, 0.960]
RND-Gold-nDCG	1	1	**0.752** [0.559, 0.868]	0.817 [0.665, 0.904]
PRI-Gold-nERR	0.869 [0.754, 0.932]	1	0.882 [0.777, 0.939]	**1**
RND-Gold-nERR	**0.908** [0.824, 0.953]	1	**0.987** [0.974, 0.993]	0.987 [0.974, 0.993]
PRI-Gold-iRBU	0.993 [0.986, 0.996]	0.993 [0.986, 0.996]	0.797 [0.632, 0.893]	0.954 [0.910, 0.977]
RND-Gold-iRBU	0.993 [0.986, 0.996]	0.993 [0.986, 0.996]	**0.941** [0.885, 0.970]	**0.993** [0.986, 0.996]

are *underestimated*. This is a way to simulate a situation where new systems (i.e., those that did not contribute to the pools) need to be evaluated using a legacy test collection.

This appendix reports on our LOTO experiments, although our sample size (i.e., the number of teams to be left out in turn) is too small for us to obtain conclusive results.

Table 16 shows the unique contributions from each WWW-4 task participant T, and the number of topic-document pairs in the lo-T qrels file. For example, SLWWW contributed as many as 2,932 unique contributions (1,467 documents to PRI-Gold topics and 1,465 documents to RND topics), so if we remove these from the original qrels file containing 10,333 topic-document pairs (See Table 2), we are left with $10,333 - 2,932 = 7,401$ pairs in the lo-SLWWW qrels file.

Table 17 addresses **RQ6.1** (*How robust to new systems are PRI-based and RND-based qrels files under a Gold environment?*) by comparing the LOTO results with the 25 PRI-Gold topics and those with the 24 RND-Gold topics, both using the Gold qrels file. For example, when the runs are ranked by mean nDCG over the 25 PRI-Gold topics, the τ between the original ranking with the Gold assessments and the the ranking based on the lo-SLWWW assessments is 0.647 (95%CI[0.400, 0.806]): the runs from Team SLWWW are substantially underestimated as this team contributed many unique relevant documents to the original pool files, as was discussed above. It can be observed that the results are inconclusive: for example, while PRI-Gold appears to be more robust than RND-Gold with nDCG, Q, and nERR in the "lo-THUIR" experiment, the opposite is true for the "lo-SLWWW" experiment. As we only have four participating teams, we cannot draw a clear conclusion from the results. Hence, to answer **RQ6.1**: *It is not clear from our Gold assessor experiments whether the document strategy has an impact on the robustness to handling new systems.* It is important to note that, in contrast to the above inconclusive result, the Bronze assessor result of Sakai et al. [41] suggested that PRI tends to outperform RND in terms of robustness to handling new systems. This may be because their Bronze-based experiment was larger in scale from several different viewpoints, as we have discussed in Section 2.3. That is, their results may be more reliable. While it would be ideal in principle to further pursue **RQ6.1** by conducting a larger Gold assessor study by running yet another task at NTCIR or elsewhere, this is practically challenging as task organisers cannot guarantee a high number of participants in advance.

Table 18 addresses **RQ6.2** (*How robust to new systems are Gold-based and Bronze-based qrels files under a PRI environment?*) by showing our LOTO experiments with Gold, BronzeW, and BronzeT qrels files, where every ranking uses the 25 PRI-Gold topics. Note that the "PRI-Gold" rows are duplicated from Table 17. Again, the trend is not very clear. For example, in the "lo-SLWWW" experiment, PRI-BronzeW is the most robust with nDCG and Q, while PRI-Gold is the most robust with nERR and iRBU; in the "lo-THUIR" experiment, PRI-Gold is the most robust with nDCG and nERR, while PRI-BronzeT is the most robust with Q and iRBU. Hence, to answer **RQ6.2**: *It is not clear from our PRI environment experiments whether the assessor type has an impact on the robustness to handling new systems.*

Table 18. Robustness in terms of Kendall's τ (with 95%CIs, $n = 18$) when compared to the system ranking before leaving out a team. The results are based on the mean scores over the 25 PRI-Gold topics; In each column (for each evaluation measure), the higher value is indicated in **bold**.

	lo-ORG	lo-KASYS	lo-SLWWW	lo-THUIR
PRI-Gold-nDCG	1	1	0.647 [0.400, 0.806]	**1**
PRI-BronzeW-nDCG	1	1	**0.804** [0.643, 0.897]	0.935 [0.874, 0.967]
PRI-BronzeT-nDCG	1	1	0.595 [0.327, 0.775]	0.948 [0.899, 0.974]
PRI-Gold-Q	1	1	0.569 [0.291, 0.758]	0.922 [0.850, 0.960]
PRI-BronzeW-Q	1	1	**0.817** [0.665, 0.904]	0.908 [0.824, 0.953]
PRI-BronzeT-Q	1	1	0.575 [0.299, 0.762]	**0.948** [0.899, 0.974]
PRI-Gold-nERR	0.869 [0.754, 0.932]	1	**0.882** [0.777, 0.939]	**1**
PRI-BronzeW-nERR	0.980 [0.960, 0.990]	0.993 [0.986, 0.996]	0.824 [0.677, 0.908]	0.993 [0.986, 0.996]
PRI-BronzeT-nERR	0.961 [0.924, 0.980]	1	0.778 [0.601, 0.882]	0.987 [0.974, 0.993]
PRI-Gold-iRBU	0.993 [0.986, 0.996]	0.993 [0.986, 0.996]	**0.797** [0.632, 0.893]	0.954 [0.910, 0.977]
PRI-BronzeW-iRBU	1	1	0.752 [0.559, 0.868]	0.974 [0.949, 0.987]
PRI-BronzeT-iRBU	1	1	0.503 [0.204, 0.716]	**1**

REFERENCES

[1] Aiman L. Al-Harbi and Mark D. Smucker. 2014. A Qualitative Exploration of Secondary Assessor Relevance Judging Behavior. In *Proceedings of the 5th Information Interaction in Context Symposium* (Regensburg, Germany). Association for Computing Machinery, 195–204.

[2] James Allan, Ben Carterette, Javed A. Aslam, Virgil Pavlu, Blagovest Dachev, and Evangelos Kanoulas. 2008. Million Query Track 2007 Overview. In *The Sixteenth Text REtrieval Conference (TREC 2007) Proceedings* (Gaithersburg, Maryland, USA). NIST.

[3] James Allan, Donna Harman, Evangelos Kanoulas, Dan Li, Christophe Van Gysel, and Ellen Voorhees. 2018. TREC Common Core Track Overview. In *The Twenty-Sixth Text REtrieval Conference (TREC 2017) Proceedings* (Gaithersburg, Maryland, USA). NIST.

[4] Omar Alonso and Stefano Mizzaro. 2012. Using Crowdsourcing for TREC Relevance Assessment. *Information Processing and Management* 48, 6 (2012), 1053–1066.

[5] 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 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Singapore, Singapore). Association for Computing Machinery, 667–674.

[6] Ben Carterette, Virgil Pavlu, Hui Fang, and Evangelos Kanoulas. 2010. Million Query Track 2009 Overview. In *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings* (Gaithersburg, Maryland, USA). NIST.

[7] Alexandra Chouldechova and David Mease. 2013. Differences in Search Engine Evaluations Between Query Owners and Non-Owners. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining* (Rome, Italy). Association for Computing Machinery, 103–112.

[8] Charles L.A. Clarke, Nick Craswell, and Ian Soboroff. 2010. Overview of the TREC 2009 Web Track. In *The Eighteenth Text REtrieval Conference (TREC 2009) Proceedings* (Gaithersburg, Maryland, USA). NIST.

[9] Cyril Cleverdon. 1970. *The effect of variation in relevance assessments in comparative experimental tests of indexing languages.* Technical Report. College of Aeronautics, Cranfield, UK.

[10] Paul Clough, Mark Sanderson, Jiayu Tang, Tim Gollins, and Amy Warner. 2012. Examining the Limits of Crowdsourcing for Relevance Assessment. *IEEE Internet Computing* 17 (2012), 32–38. Issue 4.

[11] Jacob Cohen. 1968. Weighted Kappa: Nominal Scale Agreement Provision for Scaled Disagreement or Partial Credit. *Psychological Bulletin* 70, 4 (1968), 213–220.

[12] Gordon V. Cormack, Charles L.A. Clarke, Christopher R. Palmer, and Samuel S.L. To. 1998. Passage-Based Refinement (MultiText Experiment for TREC-6). In *The Sixth Text REtrieval Conference (TREC 6)* (Gaithersburg, Maryland, USA). NIST, 303–319.

[13] Gordon V. Cormack and Maura R. Grossman. 2015. Autonomy and Reliability of Continuous Active Learning for Technology-Assisted Review. (2015). https://arxiv.org/abs/1504.06868

[14] Gordon V. Cormack, Christopher R. Palmer, and Charles L.A. Clarke. 1998. Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia). Association for Computing Machinery, 282–289.

[15] Tadele T. Damessie, J. Shane Culpepper, Jaewon Kim, and Falk Scholer. 2018. Presentation Ordering Effects on Assessor Agreement. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy). Association for Computing Machinery, 723–732.

[16] 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.

[17] Gene V. Glass, Barry McGaw, and Mary Lee Smith. 1981. *Meta-Analysis in Social Research.* Sage Publications.

[18] Donna K. Harman. 2005. The TREC Test Collections. In *TREC: Experiment and Evaluation in Information Retrieval*, Ellen M. Voorhees and Donna K. Harman (Eds.). The MIT Press, Chapter 2.

[19] 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.

[20] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Transactions on Information Systems* 20, 4 (2002), 422–446.

[21] Noriko Kando. 2004. Evaluation of Information Access Technologies at the NTCIR Workshop. In *Comparative Evaluation of Multilingual Information Access Systems (Lecture Notes in Computer Science 3237)* (Trondheim, Norway), Carol Peters, Julio Gonzalo, Martin Braschler, and Michael Kluck (Eds.). Springer, 29–43.

[22] Maurice G. Kendall. 1962. *Rank Correlation Methods (3rd Edition).* Charles Griffin and Company Limited.

[23] Kenneth A. Kinney, Scott B. Huffman, and Juting Zhai. 2008. How Evaluator Domain Expertise Affects Search Result Relevance Judgments. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management* (Napa Valley, California, USA). Association for Computing Machinery, 591–598.

[24] Mucahid Kutlu, Tyler McDonnell, Yassmine Barkallah, Tamer Elsayed, and Matthew Lease. 2018. Crowd vs. Expert: What Can Relevance Judgment Rationales Teach Us About Assessor Disagreement?. In *The 41st International ACM SIGIR Conference on Research*

and Development in Information Retrieval (Ann Arbor, MI, USA). Association for Computing Machinery, 805–814.

[25] 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.

[26] Jeffrey D. Long and Norman Cliff. 1997. Confidence Intervals for Kendall's tau. *Brit. J. Math. Statist. Psych.* 50 (1997), 31–41.

[27] 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.

[28] 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* (2018), 49–60.

[29] Tyler McDonnell, Matthew Lease, Mucahid Kutlu, and Tamer Elsayed. 2016. Why Is That Relevant? Collecting Annotator Rationales for Relevance Judgments. In *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing. 4, 1* (Phoenix, Arizona USA). AAAI, 139–148.

[30] Tetsuya Sakai. 2007. Alternatives to Bpref. In *Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Amsterdam, The Netherlands). Association for Computing Machinery, 71–78.

[31] Tetsuya Sakai. 2014. Metrics, Statistics, Tests. In *Bridging Between Information Retrieval and Databases. PROMISE 2013. Lecture Notes in Computer Science 8173*, Nicola Ferro (Ed.). 116–163.

[32] Tetsuya Sakai. 2018. Laboratory Experiments in Information Retrieval: Sample Sizes, Effect Sizes, and Statistical Power. Springer.

[33] Tetsuya Sakai. 2018. Q-measure. In *Encyclopedia of Database Systems*, Ling Liu and M. Tamer Özsu (Eds.). Springer. https://doi.org/10.1007/978-1-4899-7993-3_80616-1

[34] Tetsuya Sakai. 2019. How to Run an Evaluation Task. In *Information Retrieval Evaluation in a Changing World*, Nicola Ferro and Carol Peters (Eds.). Springer.

[35] Tetsuya Sakai. 2020. Graded Relevance. In *Evaluating Information Retrieval and Access Tasks: NTCIR's Legacy of Research Impact*, Tetsuya Sakai, Douglas W. Oard, and Noriko Kando (Eds.). Springer, 1–20.

[36] Tetsuya Sakai. 2021. On the Instability of Diminishing Return IR Measures. In *Advances in Information Retrieval. ECIR 2021. Lecture Notes in Computer Science 12656*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). 572–586.

[37] 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 IR4QA Task. In *Proceedings of the 7th NTCIR Workshop Meeting on Evaluation of Information Access Technologies: Information Retrieval, Question Answering, and Cross-Lingual Information Access* (Tokyo, Japan). National Insitute of Informatics, 77–114.

[38] Tetsuya Sakai, Sijie Tao, Zhumin Chu, Maria Maistro, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, and Yiqun Liu. 2022. Overview of the NTCIR-16 We Want Web with CENTRE (WWW-4) Task. In *Proceedings of The 16th NTCIR Conference: Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics, 234–245.

[39] Tetsuya Sakai, Sijie Tao, Maria Maistro, Zhumin Chu, Yujing Li, Nuo Chen, Nicola Ferro, Junjie Wang, Ian Soboroff, and Yiqun Liu. 2022. Corrected Evaluation Results of the NTCIR WWW-2, WWW-3, and WWW-4 English Subtasks. (2022). https://arxiv.org/abs/2210.10266

[40] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents? *ACM TOIS* 40, 4, Article 76 (2022).

[41] Tetsuya Sakai, Sijie Tao, and Zhaohao Zeng. 2022. Relevance Assessments for Web Search Evaluation: Should We Randomise or Prioritise the Pooled Documents? (CORRECTED VERSION). (2022). http://arxiv.org/abs/2211.00981

[42] Tetsuya Sakai and Peng Xiao. 2019. Randomised vs. Prioritised Pools for Relevance Assessments: Sample Size Considerations. In *Information Retrieval Technology. AIRS 2019. Lecture Notes in Computer Science 12004* (Hong Kong, China). Springer, 94–105.

[43] 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 36th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Dublin, Ireland). Association for Computing Machinery, 623–632.

[44] Milad Shokouhi, Ryen W. White, and Emine Yilmaz. 2015. Anchoring and Adjustment in Relevance Estimation. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Santiago, Chile). Association for Computing Machinery, 963–966.

[45] Eero Sormunen. 2002. Liberal relevance criteria of TREC - counting on negligible documents?. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland). Association for Computing Machinery, 324–330.

[46] 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.

[47] 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.

[48] Yuya Ubukata, Masaki Muraoka, Sijie Tao, and Tetsuya Sakai. 2022. SLWWW at the NTCIR-16 WWW-4 Task. In *Proceedings of The 16th NTCIR Conference: Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics, 243–246.

[49] Kota Usuha, Kohei Shinden, and Makoto P. Kato. 2022. KASYS at the NTCIR-16 WWW-4 Task. In *Proceedings of The 16th NTCIR Conference: Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics, 247–253.

[50] Ellen M. Voorhees. 2000. Variations in Relevance Judgments and the Measurement of Retrieval Effectiveness. *Information Processing and Management* 36 (2000), 697–716. Issue 5.

[51] Ellen M. Voorhees. 2018. On Building Fair and Reusable Test Collections using Bandit Techniques. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management* (Torino, Italy). Association for Computing Machinery, 407–416.

[52] Ellen M. Voorhees and Chris Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland). Association for Computing Machinery, 316–323.

[53] Ellen M. Voorhees, Nick Craswell, and Jimmy Lin. 2022. Too Many Relevants: Whither Cranfield Test Collections?. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain). Association for Computing Machinery, 2970–2980.

[54] Ellen M. Voorhees and Donna Harman. 2000. Overview of the Eighth Text REtrieval Conference (TREC-8). In *NIST Special Publication 500-246: The Eighth Text REtrieval Conference (TREC 8)* (Gaithersburg, Maryland, USA). NIST.

[55] Simon Wakeling, Martin Halvey, Robert Villa, and Laura Hasler. 2016. A Comparison of Primary and Secondary Relevance Judgements for Real-Life Topics. In *Proceedings of the 2016 ACM on Conference on Human Information Interaction and Retrieval* (Carrboro, North Carolina, USA). Association for Computing Machinery, 173–182.

[56] Shenghao Yang, Haitao Li, Zhumin Chu, Jingtao Zhan, Yiqun Liu, Min Zhang, and Shaoping Ma. 2022. THUIR at the NTCIR-16 WWW-4 Task. In *Proceedings of The 16th NTCIR Conference: Evaluation of Information Access Technologies* (Tokyo, Japan). National Institute of Informatics, 254–257.

[57] Emine Yilmaz and Javed A. Aslam. 2006. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management* (Arlington, Virginia, USA). Association for Computing Machinery, 102–111.

[58] Justin Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments?. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Melbourne, Australia). Association for Computing Machinery, 307–314.