How do you Test a Test?
A Multifaceted Examination of Significance Tests

Nicola Ferro
University of Padova
ferro@dei.unipd.it

Mark Sanderson
RMIT University
mark.sanderson@rmit.edu.au

ABSTRACT

We examine three statistical significance tests – a recently proposed ANOVA model and two baseline tests – using a suite of measures to determine which is better suited for offline evaluation. We apply our analysis to both the runs of a whole TREC track and also to the runs submitted by six participant groups. The former reveals test behavior in the heterogeneous settings of a large-scale offline evaluation initiative; the latter, almost overlooked in past work (to the best of our knowledge), reveals what happens in the much more restricted case of variants of a single system, i.e. the typical context in which companies and research groups operate. We find the ANOVA test strikingly consistent in large-scale settings, but worryingly inconsistent in some participant experiments. Of greater concern, the participant only experiments show one of our baseline tests (a test widely used in research) can produce a substantial number of inconsistent results. We discuss the implications of this inconsistency for possible publication bias.

CCS CONCEPTS

• **Information systems → Evaluation of retrieval results**; **Retrieval effectiveness**;

KEYWORDS

statistical significance testing; comparing tests; ANOVA; prediction

ACM Reference Format:
Nicola Ferro and Mark Sanderson. 2022. How do you Test a Test? A Multifaceted Examination of Significance Tests. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining (WSDM '22), February 21–25, 2022, Tempe, AZ, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3488560.3498406

1 INTRODUCTION

Gruson et al. [14] described how offline evaluation is a valuable correlated complement to online evaluation methods. Statistical significance testing is central to offline evaluation of search to help understand which of two alternative systems or *runs* has better performance. It is critical, therefore, to understand how stable and generalizable the significance tests are so that we can more confidently select offline outcomes. Researchers have attempted to establish, which test (e.g. Sign, Wilcoxon, t, randomization, etc.) is more suited to offline experimentation. A range of methods exist to determine which test is best, however, researchers often focus on a single property, e.g. how many significantly different system pairs are detected or how much tests agree with each other.

While a number of methods have been described in the literature that compare significance tests, we contend that few of the existing methods consider a necessary range of analyses. Therefore, here, we extend an existing method of comparison to provide a deeper understanding of the behavior of an *ANalysis Of VAriance (ANOVA)* test and the commonest significant test in *Information Retrieval (IR)* research: the t-test. We describe *multiple indicators*, each considering different angles of significance tests. Through their joint use, we obtain a more comprehensive analysis of the behaviour of a significance test and its trade-offs. We explore the use of significance across the runs of multiple retrieval systems and the runs from six single systems. While the former is the typical setting in which statistical tests are commonly examined, the latter is, to the best of our knowledge, almost completely overlooked in the literature, yet is much closer to the actual settings in which companies and research groups operate.

We choose ANOVA, because a new approach to determining the significance of a difference measured between two IR systems, or *runs*, has been proposed in recent publications. The approach shards the documents of a test collection, which significance tests exploit for greater power. Two such tests have been described in the literature [12, 50], both are claimed to better model previously overlooked interaction effects between topics and runs [12, 50], shards and runs [12], as well as topics and shards [12]. The behavior of the tests has been measured [8, 12, 50], however, the main focus of past work is an examination of the number of significant differences observed and properties of test outputs. There is some suggestion that one of the tests may produce a notable number of Type I errors (i.e. false positives) [50]; the other test has not yet been examined for this potential flaw [12]. No publication has sought to understand why the tests behave as they do, or to quantify or contextualize such errors. This paper begins to rectify this omission by asking the following research questions:

(1) Is the ANOVA-based significance test described by Ferro and Sanderson [12] superior to a commonly used existing test?
(2) Is there a benefit in comparing tests in a broader manner than has been tried in past work?

Starting with an examination of past work, the paper next describes the range of analyses used to examine the tests. The experimental setup is then described followed by a detailing of the experimental results, which are discussed in the context of past work before conclusions are drawn and future work described.

WSDM '22, February 21–25, 2022, Tempe, AZ, USA.

ACM ISBN 978-1-4503-9132-0/22/02…$15.00
https://doi.org/10.1145/3488560.3498406

2 RELATED WORK

We describe the use of significance tests in IR experimentation and review methods that determine the best for experimentation.

2.1 The use of significance tests

Significance testing in early IR experiments is rare. Fels [9] applied Fisher's exact test on a small document collection, Lesk [24] and then Salton and Lesk [32] detailed the application of the sign and t-test, Ide [18] investigated the Wilcoxon. The early demonstrations did not translate into widespread adoption of significance testing by the research community. At the time, probably the best known mention of the word "significance" came from the 5% and 10% absolute difference rules of thumb of Spärck Jones [38].

The wider use of significance tests started with Tague-Sutcliffe and Blustein's extensive application of statistical methods to TREC results [42]. Savoy [35] detailed the bootstrap test for drawing samples of topics and calculating confidence intervals (Cormack and Lynam [7] bootstrapped samples of documents), Zobel [52] described an ANOVA test that was similar to the t-test, and Smucker et al. [36] examined the use of the randomization test.

2.2 The development of new significance tests

The tests were used in a way that assumed that the main source of variability in the measurement of the effectiveness of a run was found in the topics of a test collection. However, dependent sources of variability have been determined between runs and topics, between topics and collections, and to a lesser extent between runs and collections [10].

In reaction to these dependencies, two groups created customized ANOVA models to measure effectiveness and determine significance, taking some or all of the dependencies into account. Voorhees et al. [50] used a bootstrap ANOVA approach that drew on a sample of the scores of topics measured across different systems and a document collection split into two or three shards. In a series of papers, Ferro, Kim, and Sanderson developed an ANOVA model using multiple shards [10–12]. Both research efforts claimed their models produced superior fidelity over existing significance tests but both efforts reported limited evaluation. The test from Voorhees et al. appeared to produce a number of Type I errors, but the errors were not quantified. The other test was not examined for this error [12].

This prompts the question, how good are these new tests?

2.3 Differentiating tests

Many methodologies have been employed to determine what is the most effective significance test.

2.3.1 Examining test assumptions. Early assessments of tests questioned the validity of their use. The creators of tests such as the sign, Wilcoxon, and t detailed necessary assumptions of the data that tests were applied to. Saracevic [34] stated that test collection data "*does not satisfy the rigid assumptions under which such tests are run*" (p. 13), van Rijsbergen [47, chap. 7] reiterated Saracevic's concern, but after pointing out the low number of assumptions violated for the sign test, suggested it be used "*conservatively*".

Hull [17] later questioned if assumption invalidation was of sufficient practical importance to justify significance test neglect.

2.3.2 Compare test results. Another means of differentiating tests was to compare the results of tests on a common set of system outputs, i.e. *runs*, an approach Keen [22] used to examine the Wilcoxon and sign tests. Hull's early comparisons were more extensively implemented by Zobel [52] who found that the ANOVA and t-test produced similar results to each other, the Wilcoxon, somewhat different. Smucker et al. [36] compared the p-values of the Wilcoxon, sign, randomization, and t-test, as well as the sign minimum difference test and bootstrap shift. Finding that use of many of the tests resulted in similar conclusions, Smucker et al. concluded that the randomization test was the best option for experimentation as it could be used more flexibly than other tests. In a later comparison using different topic set sizes, the same researchers concluded that the randomization and t-test were good candidates [37]. Voorhees et al. [50] visualized such comparisons.

The approaches so far do not answer the question researchers typically ask: which test finds the largest number of 'real'[1] significant differences, while producing the fewest errors? The two key errors that significance tests make are Type I (false positives) and Type II (false negatives). A series of methodologies have been employed that either explicitly or implicitly measure the balance of such errors.

2.3.3 Topic splitting. The topics of a test collection are randomly split into two Topic Sets (TSs), a run is compared with another across both TSs. The topic splitting methodology simulates a repetition of an offline experiment examining if the comparison in one TS is predictive of the same comparison in the other TS. The level of agreement across the two TSs is used as a proxy of test errors.

To the best of our knowledge, Zobel [52] introduced this technique and was the first to empirically compare significance tests applied to IR experimental results. Zobel considered the ANOVA, Wilcoxon, and t-test. Agreement across the two TSs was high for all three, but the Wilcoxon was concluded by Zobel to be the best. Voorhees and Buckley [49] used the methodology to examine the impact of TS size on evaluation consistency. Sanderson and Zobel [33] applied topic splitting to examine the sign, Wilcoxon, and t-test, concluding that the t-test was the most reliable. Sanderson and Zobel also examined the p value of a test, finding that for lower values of p, agreement across the two TSs was commoner.

The definition of agreement between the TS varied. Zobel's asked if one run was found to be significantly better than another on one set, and if the sign of the difference in runs was preserved on the other, there was agreement. Moffat et al. [25], comparing different evaluation measures (not significance tests), identified five categories of agreement or disagreement that such a comparison could result in. One form, called SSA, required the same significant improvement to be found in both sets; a more stringent version of Zobel's definition. Urbano et al. [45] created five categories of agreement. The union of two of the categories, "Success" and "Lack of power", aligns with Zobel's definition.

A flaw in the splitting methodology is that the two TS are typically disjoint, once the topics of one set are determined, the other set must contain the topics that remain [33]. As an attempt to address this, Boytsov et al. [2] used a test collection with a large number of

[1] We realize that conclusions drawn from significance tests determine whether to accept or reject the null hypothesis, but we use this terminology for convenience.

topics (30K). If one run was found to be better than another when measured on the full topic set, it was assumed the result was correct. Significance test reliability was measured on samples drawn from the 30K. A test was deemed reliable if it showed the same result between the two runs on the sample of topics as the full set. Boytsov et al. investigated test correction methods (e.g. Bonferroni), showing that without such tests, an unadjusted significance test such as the t-test was likely to return many Type I errors.[2]

2.3.4 Is it the same system? Almost all significance testing in IR use paired tests, Sakai [30] sought the best test when unpaired testing was employed. The methodology examined if a test could determine that two runs were actually from the same IR system. The author produced two runs from a single system by having it retrieve from two TS. The number of false positives was measured. Sakai examined two versions of the t-test, Student's and Welch's, finding that Student's version was better.

2.3.5 Modelling retrieval. Another approach is to simulate retrieval; the simulation can be adjusted to change an IR system to be better or worse and a significance test can be applied to see if the change can be detected. While simulation of IR systems has long been examined [6, 15, 41], it would appear that Wilbur [51] was the first to apply this modeling approach in significance, building simulated retrieval systems searching over pre-TREC test collections. Wilbur concluded that non-parametric tests (e.g. randomisation and bootstrap) were superior to parametric tests.

Parapar et al. [27] modelled a new run using a stochastic process to simulate retrieval rankings either from a rank of document scores, or just a rank of documents [26]. While the models produced a retrieval system that is on average improved by a certain percentage, the random process made some topics better, and some worse. However, the models represent a simplification of how topic variability occurs, which has been shown to be a dependent process between topics and runs [42]: one run may be better than another, improving some topics, making others worse, however the topics that are improved or made worse is in part dependent on the properties of the system and is not a stochastic process.

Urbano and Nagler [46] introduced use of copula models in part to capture the topic*run dependency. Urbano et al. [44] later manipulated the model to create a large number of simulated topics, which they used to test the power of different significance tests. They found that the permutation and t-tests worked well.

2.3.6 Measure test properties. One can examine the outputs of a test. For example Faggioli and Ferro [8], Ferro et al. [10], Ferro and Sanderson [11, 12], Voorhees et al. [50] considered their ANOVA based tests superior if the number of significant tests found by them was larger than that found by alternative tests. Voorhees et al. [50] examined the size of confidence intervals resulting from their test and alternatives, preferring tests that produced smaller intervals. Ferro et al. [10] examined un-modelled error levels. Ferro and Sanderson [12] analytically showed that an ANOVA model comprising the topic, run, and shard factors as well as their interaction is the only one which leads to an estimation of the run factor independent from any missing values caused by shards without a relevant document. To the best of our knowledge, this is one of the very few formal demonstrations of the properties of system comparison.

2.4 Publication bias

It has long been recognized that a side effect of a researcher obtaining significance in an experimental result is that the result appears to have a greater chance of publication [39], the effect is known as *publication bias*. Sakai [29] extensively surveyed experimental IR papers showing the majority reported significance. Carterette [5] graphed the distribution of p-values reported in thirty IR papers, finding no evidence of publication bias in those papers.

3 APPROACH

In past work, topic splitting and modeling retrieval were the main approaches to differentiating significance tests. As we will examine one of the new class of significance tests that consider dependencies between topics, runs, and collection shards, we are unable to select from the modeling approaches as currently none consider these dependencies. Being an approach derived from data, the topic splitting methodology will encompass our considered dependencies. However, the methodology must be treated with care. As pointed out by Parapar et al. [26], topic splitting "*works with ... small splits of queries*" and measures consistency, not correctness, a "*test might be consistently rejecting a null hypothesis that is true or, conversely, it might be consistently accepting a null hypothesis that is actually false*". Consequently, we use a test collection with a large number of runs, a document collection and that can be sharded, and crucially, many topics. We also introduce fake models that consistently reject the null hypothesis in order to understand its behavior.

3.1 Use Case: The ANOVA Models

We consider the following ANOVA models:

$$y_{ij} = \mu_{\cdot\cdot} + \alpha_j + \varepsilon_{ij} \tag{MD1}$$

$$y_{ij} = \mu_{\cdot\cdot} + \tau_i + \alpha_j + \varepsilon_{ij} \tag{MD2}$$

$$y_{ijk} = \mu_{\cdot\cdot\cdot} + \tau_i + \alpha_j + \beta_k + (\tau\alpha)_{ij} + (\tau\beta)_{ik} + (\alpha\beta)_{jk} + \varepsilon_{ijk} \tag{MD3}$$

where:

- y_{ij} is the performance score of the j-th run on the i-th topic, when we use the whole corpus;
- y_{ijk} is the performance score of the j-th run on the i-th topic for the k-th shard, when we use shards;
- $\mu_{\cdot\cdot}$ (or $\mu_{\cdot\cdot\cdot}$ in the case of shards) is the grand mean;
- τ_i is the effect of the i-th topic;
- α_j is the effect of the j-th run;
- β_k is the effect of the k-th shard;
- $(\tau\alpha)_{ij}$ is the interaction between topics and runs;
- $(\tau\beta)_{ik}$ is the interaction between topics and shards;
- $(\alpha\beta)_{jk}$ is the interaction between runs and shards;
- ε_{ij} (or ε_{ijk} in the case of shards) is the error committed by the model in predicting y.

(MD1) considers just the run effect on the whole corpus. When comparing two runs, (MD1) is equivalent to an unpaired t-test, since it does not consider that the same topics have been experimented

[2] See also Carterette [4] who examined multiple comparison corrections.

with both runs. We consider this test as a lower bound of our experiments.

(MD2) considers both the topic and the run effects on the whole corpus also. When comparing two runs, (MD2) is equivalent to a paired t-test, since it considers that the same topics have been experimented with both runs. (MD2) was developed by Tague-Sutcliffe and Blustein [42] and later used by Banks et al. [1]; it also corresponds to model of equation (2) of Voorhees et al. [50] and to (MD2) of Ferro and Sanderson [12]. As Sakai [29]'s extensive survey found the t-test to be the most widely used significance test in IR research papers, we consider this model to be our baseline.

(MD3) is more complex: it requires the corpus to be sharded and considers topic, run, and shard effects, together with all their interactions. It corresponds to model (MD6) of Ferro and Sanderson [12]. Voorhees et al. did not consider this model.

The system effect factor α_j is given by the marginal mean of the j-th factor minus the grand mean. For (MD1) and (MD2) the marginal mean j-th factor is $\hat{\mu}_{\cdot j} = \frac{1}{T}\sum_{i=1}^{T} y_{ij}$. In the case of (MD3), the marginal mean j-th factor is $\hat{\mu}_{\cdot j \cdot} = \frac{1}{T \cdot S}\sum_{i=1}^{T}\sum_{k=1}^{S} y_{ijk}$, which is the average of system performance across both topics and shards.

We create two "*Fake* ANOVA" models, which declare all run pairs measured differently to be significant. They are obtained by computing scores and Tukey *Honestly Significant Difference (HSD)* correction as usual but then always setting $p = 0.0$. The models provide upper bounds for our analysis. The first, MD2f, corresponds to both eq (MD1) and (MD2) as the estimated system factor is the same in both models; MD3f corresponds to eq (MD3).

3.2 Analysis Methods

We use *topic splitting*, sampling without replacement to form two topics sets – TS1 and TS2 – of varying sizes. We repeated this procedure 100 times (as described in Section 4). We computed the ANOVA models on each TS and consider the level of agreement on the two sets according to several measures, described next. We conduct the following analyses:

- Drawing from past work, [33, 45, 52], we use the same model on both TSs. This allows us to understand how much the conclusions we draw from one experiment, hold for another.
- We use a different model on each TS, e.g. compare (MD2) on TS1 against (MD3) on TS2. The analysis reveals the extent that conclusions drawn from one model hold for another.
- We restrict analysis to the runs submitted by single participants.

3.3 Analysis Measures

We considered different measures to compare the two TSs.

Jaccard Similarity. We measure the Jaccard similarity [19] between the sets of significantly different run pairs on two TSs. The score tells us the closeness of the two sets.

$$J = \frac{|TS1 \cap TS2|}{|TS1 \cup TS2|}$$

Overlap Coefficient. We measure overlap [40] between the sets of significantly different run pairs on the two TSs. The score quantifies how much one set is a subset of the other.

$$O = \frac{|TS1 \cap TS2|}{\min(|TS1|, |TS2|)}$$

The Jaccard and Overlap coefficients provide a high level overview of similarity of significance between run pairs, but we need measures to give us greater fidelity, we utilize the following:

- **Active Agreement (AA)**: on both TS1 and TS2, either $S1 >> S2$ or $S2 >> S1$. A consistent outcome on what is significantly different. The larger the AA is, the better. It corresponds to SSA [25], Success [45], and Active Agreements [8] in past work.
- **Active Disagreements (AD)**: $S1 >> S2$ on TS1 but $S2 >> S1$ on TS2; or $S2 >> S1$ on TS1 but $S1 >> S2$ on TS2. The worst outcome, since there is inconsistent agreement on which run is significantly better. The two TSs lead to conflicting conclusions. The smaller AD is, the better. It corresponds to SSD [25], Major Error [45], and Active Disagrements [8] in past work. Note, in the case of the fake ANOVAs, MDf2 and MDf3, the AD count is similar to the notion of swaps [49], since it detects how many times the difference in the marginal means of the system factor disagree on the two TSs.
- **Mixed Agreement (MA)**: $S1 >> S2$ on TS1 but $S1 > S2$ on TS2; or $S2 >> S1$ on TS1 but $S2 > S1$ on TS2; or $S1 > S2$ on TS1 but $S1 >> S2$ on TS2; or $S2 > S1$ on TS1 but $S2 >> S1$ on TS2; The smaller, the better, since while the order of the two runs is the same, it indicates a situation where a model is not able to confirm its conclusions on both TSs. It corresponds to SN + NS [25], Lack of Power [45], and Passive Disagrements [8] in past work.
- **Mixed Disagreement (MD)**: $S1 >> S2$ on TS1 but $S2 > S1$ on TS2; or $S2 >> S1$ on TS1 but $S1 > S2$ on TS2; or $S1 > S2$ on TS1 but $S2 >> S1$ on TS2; or $S2 > S1$ on TS1 but $S1 >> S2$ on TS2. The smaller, the better, since it indicates a situation where a model is not able to confirm its conclusions on both TSs and the order of the two systems is the opposite; therefore, it is a more severe issue than MA but less than AD. It corresponds to Minor Error [45].

Kendall's τ correlation. We consider the Kendall's τ correlation [23] between the *Rankings of Systems (RoS)* on the two TSs, indicating how similarly runs were ranked on the two sets. Note that τ is independent of the ANOVA models when we use the whole corpus – i.e. we have the same τ for (MD1) and (MD2) – however, the run rank is different with shards, therefore, we compute a different τ result for (MD3).

4 EXPERIMENTAL SETUP

We detail the collection, measures, and methods of our experiments.

Collection. We used the TREC 13 (T13) robust track [48], which contains 249 topics, 110 runs (retrieving 1,000 documents per topic), and a corpus of 528K news documents (disks 4&5 of the TIPSTER collection minus Congressional Record). Relevance judgments are based on a pool depth of 100-125 documents. The judgements are multi-graded, which we mapped to binary where everything above not relevant is considered relevant.

Measures. We used Average Precision [3], Precision at 10, and *Normalized Discounted Cumulated Gain (nDCG)* [20]. Finding little difference between the three measures, we report nDCG only.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	54.3	108.4	595.2	1023.2	1815.3
MD2	336.3	760.5	1484.6	2116.2	3119.1
MD3	2517.6	3008.9	3844.1	4389.6	4920.7

Table 1: Number of significant across the 5,995 T13 run pairs for the three models on TS1.

Shards. To compute (MD3), we randomly sampled shards of documents drawn from the T13 collection without replacement, i.e. we randomly partitioned the collection. The sampling was repeated once. We examined shards of size 2, 3, 5, but found insubstantial differences between results using the different shards. Measuring with a larger number of shards was computationally intensive, therefore we selected shards of size three.

Topic Sets. We sampled topics without replacement to form two separate TSs. We considered splits of 2%, 4%, 10%, 20%, and 50% of the whole T13 TS, resulting in two splits containing, respectively, 5, 10, 25, 50, and 125[3] topics. For each split, we repeated the topic sampling 100 times and took the arithmetic mean, resulting in counts having non-integer values.

Significance Level and Multiple Comparison. We set the significance level α to 0.05. In order to control for the increased Type-I errors due to the multiple comparisons between all the possible pairs [13, 31], we adopt the Tukey HSD correction [16, 43] which ensures a *Family-wise Error Rate (FWER)* at the significance level α.

Participants. Almost all topic splitting papers of the past consider runs from a wide range of participants. However, a great many researchers use significance tests only to determine differences between the runs of a single IR system. For the participant analyses we considered the following participants and runs from T13: FUB, the 3rd top participant in terms of *Average Precision (AP)*; UoG, the 4th top participant in terms of AP. Note, we also examined, but do not report here the runs from participants, VTU, PIRC, ICL, and HUM; similar results were found.

Reproducibility. Our source code is available at the following git repository: https://bitbucket.org/frrncl/wsdm2022-fs/src/master/.

5 ANALYSIS

The results in Table 1 echo past papers by tabulating the number of significant found for the ANOVA models. We see MD3 produces substantially more significant differences than MD1 and MD2. For small topic sets, MD3 produces 1 − 2 orders of magnitude more differences than MD2 and MD1, respectively. The count, however, says nothing about the reliability of the three models. For that, we need to consider both TSs using our chosen analyses.

5.1 Jaccard and Overlap

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.02	0.10	0.48	0.77	0.84
MD2	0.20	0.47	0.69	0.75	0.84
MD3	0.45	0.56	0.70	0.79	0.87

Table 2: Jaccard across the three models on T13.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.98	1.00	1.00	1.00	1.00
MD2	0.95	0.98	0.99	1.00	1.00

Table 3: Overlap of MD3 with MD1/2 on T13 run pairs.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	2.5	21.9	365.2	903.5	1656.9
MD2	118.5	479.0	1188.12	1802.9	2846.6
MD3	1518.7	2107.8	3147.1	3855.7	4556.6
%MD1	-	-	*762%*	*327%*	*175%*
%MD2	*1,181%*	*340%*	*165%*	*114%*	*60%*

Table 4: AA on 5,995 T13 run pairs. The % relative difference between MD3 and MD1/2 is also shown. We omit the relative differences with a small divisor.

Table 2 shows the Jaccard measure between sets of significant pairs across TS1 and TS2. We see that the Jaccard for MD3 is consistently higher than for MD2 and MD1: if a significant difference is found between runs on one TS, that difference is reflected in the other TS more often for MD3 than for MD2 or MD1. We also see that as the TS size increases, the Jaccard increases and the difference between the models reduces to a point where for topic sets of 50 and 125 topics there is almost no difference. This indicates that the typical topic set sizes found in experimentation, i.e. 50 or more topics, results in comparable levels of consistency across the models. However, should one have to use smaller topics sets, MD3 may be a preferable option.

Overlap measures if the differences found by one test are the same as those found by another. We measured the Overlap between the significant pairs of MD1 and MD2 with the pairs of MD3. In Table 3, we see Overlap is at or close to 1.0 throughout: the significant pairs found by MD1/MD2 are near perfect subsets of the pairs found by MD3.

5.2 Consistency vs inconsistency: a potential for publication bias

Measuring a count of active agreements (AA) across the TSs indicates how often a model finds significance consistently. We see in Table 4 that AA across the TSs is larger for MD3 than for MD2 or for MD1. Even for small TSs, the count of AA is high for MD3: e.g. for 5 topics, the 1518.7 run pairs found in active agreement is 25.3% of all the 5,995 pairs examined. In contrast, for MD1 and MD2, there are 0.4% and 2.0% pairs found respectively. Table 4 also shows the % difference of AA for MD3 compared to the other two models. Across the TSs the difference is at least 60% higher than MD2 for large TSs and substantially higher for smaller TSs (i.e. 165% higher for TSs of 25, 340% higher for 10 topics). If we compare TS sizes across the models: we see that for MD2, the count of AA at 125 topics is (roughly) similar to the AA count from MD3 for 25 topics: 2846.6 vs 3147.1, respectively.

[3] As there are 249 topics in T13, the 50%-50% split is actually 125-124.

Considering the AD measure (Table 5), we see that unlike MD1 and MD2, MD3 results in some inconsistent behavior: on one TS a run is found to be significantly better, but on the other, that run is significantly worse. The number of ADs is small compared to the number of AAs: 2.4% for the smallest topic set considered, < 1% for all other topic sets, nevertheless inconsistent results are not desired. We use Kendall's τ to measure the similarity of the RoS. For small TSs τ is low, indicating that when TS1 and TS2 are small, the two TSs rank runs differently.

In Table 6 we see that while AD is 0 for MD1 and MD2, the two models show other forms of inconsistent behavior in MA and MD. We contrast counts of inconsistencies (AD, MA, MD) with counts of consistencies (AA) by considering the risk of publication bias when using significance tests. An AA indicates little to no chance of publication bias as regardless of TS, the same significant test result was observed. With the other three counts, on different TSs different significant test outcomes occur.

We measure *Bias* as the likelihood of a researcher publishing a significant result when in fact a significance test on a different TS would have produced either no significance (MA, MD) or a significant result in the opposite direction (AD). We calculate Bias using the following fraction.

$$Bias = 1 - \frac{AA}{AA + AD + \frac{MA}{2} + \frac{MD}{2}}$$

Note, we half the count of MA and MD, as for only one of the two TSs in those counts was significance observed. We compute one minus the fraction to focus on errors, the values shown in the tables are %s.

In Table 6, for MD2, 9.0% of significant differences were inconsistently measured. While MD3 shows a higher count of AD, MA, and MD than the other two models, the inconsistencies represent a smaller fraction of pairs found with a significant difference (7.3%). Overall, we see that there is little difference between the three models in terms of potential publication bias. The main distinction between the models is in the sum of pairs that have at least one significant difference.

We also include the 'fake' ANOVA models, MD2F and MD3F that simply assume a higher score is the same as significance, a practise commonly seen in IR papers of the last century. The Bias for the fake models is lower than for MD1, MD2, and MD3, though the fake models result in the highest count of AD, arguably the worst kind of inconsistency. Note, that the AD values of MD3f are slightly lower than MD2f. This difference was observed across a number of experiments that we ran. Recall that MD3 based measures score runs across both topics and shards instead of the conventional measurement across topics alone (used by MD1 and MD2). It would appear that regardless of significance testing, there is a slight measurement advantage to using shards. However, note also that we have not conducted sufficient testing to demonstrate this effect conclusively.

Drawing all the results in this sub-section together, we find that MD3 is at the very least as consistent as MD1 and MD2 at determining significant differences between runs and is capable of finding substantially more differences than the other two tests, particularly for small TSs. All test produce similar fractions of inconsistent counts of which run is significantly better, however the types of inconsistency between the models are different.

Ts (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.0 [0.5]	0.0 [0.6]	0.0 [0.7]	0.0 [0.8]	0.0 [0.9]
MD2	0.0 [0.5]	0.0 [0.6]	0.0 [0.7]	0.0 [0.8]	0.0 [0.9]
MD3	35.9 [0.5]	21.4 [0.6]	11.6 [0.7]	8.9 [0.8]	5.8 [0.9]

Table 5: AD [and τ] on 5,995 T13 run pairs.

Topics	AA	AD	MA	MD	Bias (%)
MD1	1656.9	0.0	329.2	0.0	9.0%
MD2	2846.6	0.0	559.6	0.7	9.0%
MD3	4556.6	5.8	591.9	114.2	7.3%
MD2f	5620.9	373.7	0.0	0.0	6.2%
MD3f	5633.7	360.3	0.0	0.0	6.0%

Table 6: Consistent and inconsistent agreements of significance across the 5,995 T13 run pairs for TS of 125.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.00	0.00	0.00	0.00	0.00
MD2	0.00	0.01	0.11	0.18	0.62
MD3	0.24	0.34	0.53	0.64	0.81

Table 7: Jaccard on FUB run pairs.

5.3 Participant Model Analysis

A great many examinations of significance tests are conducted on the runs submitted to a large evaluation conference, such as TREC, CLEF, or NTCIR. Most research uses significance tests to compare runs generated from one retrieval system. We, therefore, examined how the models performed on the runs of six participant groups. For reasons of space in the paper, we show only two of those participants that illustrate different aspects of the comparisons and of our analysis: University of Glasgow (UoG) and Fondazione Ugo Bordoni (FUB). Both groups submitted ten runs, which leads to 45 pairs compared. Results in the four omitted participant runs were similar to the results from the FUB runs.

The FUB runs, shown in Table 7, show different behavior of the models compared to the analysis on all the runs of T13 (Table 2). The MD1 and MD2 models show low Jaccard similarly between the TSs, apart from MD2 at 125 topics (0.62). For 125 topics, the MD3 model shows a Jaccard similarity comparable to the one measured across all T13 runs, but for smaller TSs the similarity is less. Unlike the similarities in Table 2, there is no similarity of Jaccard values across the models for large TSs.

For the results in Tables 8, and 9, the measure of AA, particularly for larger topic sets show a high value of AA relative to the 45 FUB run pairs. Examining the size of AA for MD2 with a 125 topics (12.31), we see that MD3 results in a similar AA with between 10 and 25 topics (9.18 and 17.53). The AD count for MD3, however, is different from above, see Table 5. Even for large TSs, AD is found in 2% of pairs for MD3, even though the Kendall's τ between the RoS is relatively high.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.00	0.00	0.00	0.00	0.00
MD2	0.00	0.03	0.53	1.75	12.31
MD3	5.50	9.18	17.53	23.17	31.93
%MD2	-	-	*3,208%*	*1,224%*	*159%*

Table 8: AA on 45 FUB run pairs. The % relative difference between MD3 and MD1/2 is also shown. We omit the relative differences with a small divisor.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.0 [0.2]	0.0 [0.4]	0.0 [0.6]	0.0 [0.7]	0.0 [0.8]
MD2	0.0 [0.2]	0.0 [0.4]	0.0 [0.6]	0.0 [0.7]	0.0 [0.8]
MD3	1.7 [0.3]	1.8 [0.4]	1.1 [0.6]	1.3 [0.7]	0.9 [0.8]

Table 9: AD [and τ] on 45 FUB run pairs.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.00	0.00	0.00	0.00	0.00
MD2	0.00	0.00	0.00	0.00	0.00
MD3	0.31	0.30	0.35	0.38	0.55

Table 10: Jaccard on 45 UoG run pairs.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.00	0.00	0.00	0.00	0.00
MD2	0.00	0.00	0.00	0.00	0.00
MD3	5.04	5.53	6.95	8.96	16.40

Table 11: AA on 45 UoG run pairs.

Considering the UoG runs, in Tables 10, 11, and 12, we see a different pattern: the Jaccard similarity, AA, and AD are low and do not change substantially across different TS sizes. We also see that AD is a notable fraction of AA measures across the TSs. Here MD3 is failing. Examining the Kendall's τ (Table 12), we see that τ is low with little to no correlation between the RoS. This indicates that the topics in the two TSs are ranking runs differently. Note, the minimal number of significant differences between the UoG runs is most likely due to a similarity of the runs to each other.

For the examination of consistent and inconsistent significant pairs (Tables 13 and 14), we see that while MD2 returns no AD counts, the count of MA and MD are a notable fraction of the 45 pairs compared across the two participant groups. For MD2 on FUB, bias is 24.3%; on UoG, all the significant pairs of this widely used test are inconsistent. For MD3, bias is smaller, but there are many inconsistencies, including some ADs.

From this analysis of participants, we see that the behavior of the significance tests is quite different from the behavior seen above. All models do not perform as well as on all the runs. The MD1 and MD2 models find few significant differences, but also find a notable number of inconsistent significant differences. The MD3 model is similarly substantially less effective on participant only runs.

Tpcs (%)	5 (2%)	10 (4%)	25 (10%)	50 (20%)	125 (50%)
MD1	0.0 [0.1]	0.0 [0.1]	0.0 [0.2]	0.0 [0.3]	0.0 [0.5]
MD2	0.0 [0.1]	0.0 [0.1]	0.0 [0.2]	0.0 [0.3]	0.0 [0.5]
MD3	4.0 [0.1]	3.4 [0.1]	3.4 [0.2]	2.8 [0.3]	2.0 [0.5]

Table 12: AD [and τ] on 45 UoG run pairs.

Topics	AA	AD	MA	MD	Bias (%)
MD1	0.0	0.0	0.0	0.0	0.0%
MD2	12.3	0.0	7.8	0.1	24.3%
MD3	31.9	0.9	5.1	2.9	13.3%

Table 13: Consistent and inconsistent agreements of significance across the 45 T13 FUB run pairs for TS of 125.

Topics	AA	AD	MA	MD	Bias (%)
MD1	0.0	0.0	0.0	0.0	0.0%
MD2	0.0	0.0	1.5	0.5	100.0%
MD3	16.4	2.0	9.9	5.1	36.7%

Table 14: Consistent and inconsistent agreements of significance across the 45 T13 UoG run pairs for TS of 125.

6 DISCUSSION

We consider the results from two perspectives: how does MD3 compare to other significance tests as examined in past work; and what is the value of the analysis method used in this paper?

6.1 Comparing MD3

It is not clear if MD3 is a more effective significance test than MD2 (i.e. the t-test). The superiority of MD3 over MD2 was claimed by Ferro and Sanderson [12] based on a count of significance tests. The researchers did not consider a wider set of measures and used fewer topics: fifty compared to the 249 here. The comparison of the models showed MD3 produces substantially more consistent results than MD2. Apart from the examination of bootstrap ANOVA [50], such levels of consistency have not been seen in past work [33, 36, 52] where all showed relatively small differences between tests. For example, Sanderson and Zobel [33] stated that "*sign and Wilcoxon measured at* $0.01 < p \leq 0.05$ *– were not as accurate as the t-test: producing respectively 17%-8% and 16%-10% error rates compared to 13%-7% error for the t-test.*". Here, in Table 4 MD3 was found to be in relative terms at least 60% more accurate than MD2.

However, the work here also showed a number of inconsistencies. While Voorhees et al. showed through visualisation (but not quantification), inconsistent results of their test, here such inconsistencies were measured. While the count of AD in Table 5 were relatively small, for the participant only runs (Tables 9 and 12) the AD count was high. Use of Kendall's τ provided a reason why AD was high, the two TSs produced notably different RoS. Examination of participant only runs when differentiating difference significance tests has been little examined in past work. Voorhees et al. [50] showed differences in confidence intervals for tests in participant only runs, but did not measure test consistency or inconsistency.

A recent study examining MD3 and bootstrap ANOVA concluded that MD3 was not as powerful, but was more stable [8]. On the question of inconsistency, Voorhees et al. [50] did show some inconsistent aspect of bootstrap ANOVA. In this paper we quantify the inconsistencies of MD3 and also of MD2 and MD1. We show that the inconsistencies are different across the models. This comparison of inconsistencies has not been shown before.

6.2 The value of the analysis

We contend the suite of analyses used in this paper provides a detailed understanding of MD3 and of MD1 and MD2. In general, many past papers that have compared significance tests to each other focused on one means of analysis.

- Smucker et al. [36] principally compared the outputs of different tests with each other, later examining the failure rates of a subset of tests.
- Sanderson and Zobel [33] used a simplified TS comparison method that failed to measure the degree of significant agreements across the sets.
- While Urbano et al. [45] detailed a range of measures, the recommendation of which test to use was ultimately driven by a calculation of global error.
- Voorhees et al. [50] compared the output of tests with each other, counting pairs of runs that were found to be significantly different, and visualizing the results.
- Past work – cited by Carterette [5] – has shown potential publication bias in researchers cherry picking results for publications. Carterette's conclusions from his work (analyzing distributions of p-values extracted from around thirty papers) suggests no evidence of such biases in the IR literature he examined. While such results are encouraging, we suspect that they are not the final word on the matter. The potential for inconsistencies identified in our work could point to an unintentional publication bias where a researcher obtains significance and publishes in good faith, but had they used a different TS (without significance), they would not have published. Although we have not tested this, we speculate that such a bias could still result in the distributions shown in Carterette's work.

We combined versions these approaches to produce, we would argue, a more complete understanding. Using more analyses has not only given us a stronger understanding of the differences between MD1, MD2, and MD3, it has also helped us understand in more detail, when tests succeed and also why and when they fail. Taking the novel step of including a 'fake ANOVA' model enabled us to see upper bounds of consistency measures, as well as highlighting the importance of understanding how a test fails across a range of error measures, as shown in Table 6.

One minor benefit of the 'fake ANOVA' models is that they showed that simply measuring effectiveness on a sharded collection resulted in slightly more consistent measurement even without a significance test, a result that has not been shown in past work.

Just examining the consistency of the MD1, MD2 and MD3 models, we see that they are similar for large TSs, not a result that we believe has been shown before. We also note the change in consistency under different topic sizes. The consistency and success of MD1 and MD2 at finding significant differences reduced substantially as size reduced. We were also able to compare the TS size where MD3 and MD2 find similar levels of AA: a size five TS with MD3 can find more consistent significant differences than MD2 with a TS with 125 topics. This style of comparison, we do not think has been shown in past work before.

With the exception of Voorhees et al.'s partial analysis (noted above), all the publications based conclusions on a large number of runs measured across multiple systems. Here, we examined the consistency and inconsistency of the three models on participant only runs. We found that both MD3 and MD2 produced an unexpected number of inconsistent pairs. To the best of our knowledge, this result has not been shown before.

7 CONCLUSIONS AND FUTURE WORK

We asked the following research questions:

1. Is the ANOVA-based significance test described by Ferro and Sanderson [12] superior to a commonly used existing test?
2. Is there a benefit in comparing tests in a broader manner than has been tried in past work?

Through widespread comparison, we show that the ANOVA MD3 is a highly consistent test compared to the well used t-test (as represented by MD2), however, a small but notable number of inconsistent results arise from use of MD3. The levels of inconsistency rise when considering participant only runs. The result suggests that this recent innovation in significance that exploits collection shards needs to be adapted to reduce the number of inconsistent results, particularly when distinguishing between participant runs.

From the discussion above, we contend that the suite of analyses used here provide a more detailed comparison of MD3 against MD1 and MD2. The analyses also help 'debug' the lack of significance found in different experiments. We contend that such a comparison has not been shown in past work. The results also show a notable number of inconsistencies in a widely used significance test, MD2 (which is equivalent to the t-test) a potential source of publication bias.

For other future work, we will examine further comparison methods, including active and passive agreements between the models across different TS. While the work here has concentrated on ANOVA based significance tests, the analyses could quite easily be applied to tests used in past work, such as the Wilcoxon, sign or randomization. We will also consider the important aspect of computational effort required to calculate significance. As has been alluded to earlier (as well as in Voorhees et al. [50]) substantial compute time is needed in order to calculate the ANOVA models based on shards. While the tests are potentially more accurate that the common t-test baseline, determining ways that they can be calculated more efficiently would be of benefit.

Acknowledgments. This research was conducted in part by the ARC Centre of Excellence for Automated Decision-Making and Society, and funded partially by the Australian Government through the Australian Research Council (CE200100005).

REFERENCES

[1] D. Banks, P. Over, and N.-F. Zhang. 1999. Blind Men and Elephants: Six Approaches to TREC data. *Information Retrieval* 1, 1-2 (May 1999), 7–34.

[2] L. Boytsov, A. Belova, and P. Westfall. 2013. Deciding on an Adjustment for Multiplicity in IR Experiments, See [21], 403–412.

[3] C. Buckley and E. M. Voorhees. 2005. Retrieval System Evaluation. In *TREC. Experiment and Evaluation in Information Retrieval*, D. K. Harman and E. M. Voorhees (Eds.). MIT Press, Cambridge (MA), USA, 53–78.

[4] B. A. Carterette. 2012. Multiple Testing in Statistical Analysis of Systems-Based Information Retrieval Experiments. *ACM Transactions on Information Systems (TOIS)* 30, 1 (2012), 4:1–4:34.

[5] B. A. Carterette. 2017. But Is It Statistically Significant? Statistical Significance in IR Research, 1995-2014. *Proc. 40th Annual International ACM SIGIR (SIGIR 2017)*. ACM Press, New York, USA, 1125–1128.

[6] Michael D Cooper. 1973. A simulation model of an information retrieval system. *Information Storage and Retrieval* 9, 1 (1973), 13–32.

[7] G. V. Cormack and T. R. Lynam. 2006. Statistical Precision of Information Retrieval Evaluation. In *Proc. 29th Annual International ACM SIGIR (SIGIR 2006)*. ACM Press, New York, USA, 533–540.

[8] G. Faggioli and N. Ferro. 2021. System Effect Estimation by Sharding: A Comparison between ANOVA Approaches to Detect Significant Differences. In *Advances in Information Retrieval. Proc. 43rd European Conference on IR Research (ECIR 2021)*. Lecture Notes in Computer Science (LNCS), Springer, Heidelberg, Germany.

[9] E. M. Fels. 1963. Evaluation of the Performance of an Information-Retrieval System by Modified Mooers Plan. *American Documentation (pre-1986)* 14, 1 (01 1963), 28. Name - University of Pittsburgh; Aslib; Copyright - Copyright Wiley Periodicals Inc. Jan 1963; Last updated - 2019-11-23.

[10] N. Ferro, Y. Kim, and M. Sanderson. 2019. Using Collection Shards to Study Retrieval Performance Effect Sizes. *ACM Transactions on Information Systems (TOIS)* 37, 3 (May 2019), 30:1–30:40.

[11] N. Ferro and M. Sanderson. 2017. Sub-corpora Impact on System Effectiveness. In *Proc. 40th Annual International ACM SIGIR (SIGIR 2017)*. ACM Press, New York, USA, 901–904.

[12] N. Ferro and M. Sanderson. 2019. Improving the Accuracy of System Performance Estimation by Using Shards, See [28], 805–814.

[13] N. Fuhr. 2017. Some Common Mistakes In IR Evaluation, And How They Can Be Avoided. *SIGIR Forum* 51, 3 (December 2017), 32–41.

[14] A. Gruson, P. Chandar, C. Charbuillet, J. McInerney, S. Hansen, D. Tardieu, and B. Carterette. 2019. Offline evaluation to make decisions about playlist recommendation algorithms. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*. 420–428.

[15] Michael D Heine. 1981. Simulation, and simulation experiments. In *Information retrieval experiment*, K. Spärck Jones (Ed.). Butterworths London, 179–198.

[16] Y. Hochberg and A. C. Tamhane. 1987. *Multiple Comparison Procedures*. John Wiley & Sons, USA.

[17] D. A. Hull. 1993. Using Statistical Testing in the Evaluation of Retrieval Experiments. In *Proc. 16th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1993)*. ACM Press, New York, USA, 329–338.

[18] E. Ide. 1968. New Experiments in Relevance Feedback. In *Report ISR-14 to the National Science Foundation*. Cornell University, Department of Computer Science.

[19] P. Jaccard. 1901. Étude comparative de la distribution florale dans une portion des Alpes et du Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles* 37, 142 (January 1901), 547–579.

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

[21] G. J. F. Jones, P. Sheridan, D. Kelly, M. de Rijke, and T. Sakai (Eds.). 2013. *Proc. 36th Annual International ACM SIGIR (SIGIR 2013)*. ACM Press, New York, USA.

[22] E Michael Keen. 1992. Presenting results of experimental retrieval comparisons. *Information Processing & Management* 28, 4 (1992), 491–502.

[23] M. G. Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1/2 (June 1938), 81–93.

[24] M. E. Lesk. 1966. SIG - The Significance Programs for Testing the Evaluation Output. In *Report ISR-12 to the National Science Foundation*. Cornell University, Department of Computer Science.

[25] A. Moffat, F. Scholer, and P. Thomas. 2012. Models and Metrics: IR Evaluation as a User Process. In *Proc. 17th Australasian Document Computing Symposium (ADCS 2012)*. ACM Press, New York, USA, 47–54.

[26] J. Parapar, D. E Losada, and Á. Barreiro. 2021. Testing the tests: simulation of rankings to compare statistical significance tests in information retrieval evaluation. In *Proceedings of the 36th Annual ACM Symposium on Applied Computing*. 655–664.

[27] J. Parapar, D. E. Losada, M. A. Presedo-Quindimil, and A. Barreiro. 2020. Using Score Distributions to Compare Statistical Significance Tests for Information Retrieval Evaluation. *Journal of the American Society for Information Science and Technology (JASIST)* 71, 1 (January 2020), 98–113.

[28] B. Piwowarski, M. Chevalier, E. Gaussier, Y. Maarek, J.-Y. Nie, and F. Scholer (Eds.). 2019. *Proc. 42nd Annual International ACM SIGIR (SIGIR 2019)*. ACM Press, New York, USA.

[29] Tetsuya Sakai. 2016. Statistical significance, power, and sample sizes: A systematic review of SIGIR and TOIS, 2006-2015. In *Proceedings of the 39th International ACM SIGIR* . 5–14.

[30] T. Sakai. 2016. Two Sample T-tests for IR Evaluation: Student or Welch?. In *Proc. 39th Annual International ACM SIGIR (SIGIR 2016)*. ACM Press, New York, USA, 1045–1048.

[31] T. Sakai. 2020. On Fuhr's Guideline for IR Evaluation. *SIGIR Forum* 54, 1 (June 2020), p14:1–p14:8.

[32] G. Salton and M. E. Lesk. 1968. Computer Evaluation of Indexing and Text Processing. *Journal of the ACM (JACM)* 15, 1 (January 1968), 8–36.

[33] M. Sanderson and J. Zobel. 2005. Information Retrieval System Evaluation: Effort, Sensitivity, and Reliability. In *Proc. 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2005)*. ACM Press, New York, USA, 162–169.

[34] Tefko Saracevic. 1968. *Comparative Systems Laboratory Final Technical Report, An Inquiry into Testing of Information Retrieval Systems Part II: Analysis of Results*. Technical Report. Case Western Reserve University.

[35] J. Savoy. 1997. Statistical inference in retrieval effectiveness evaluation. *Information Processing and Management* 33, 4 (1997), 495–512.

[36] M. D. Smucker, J. Allan, and B. A. Carterette. 2007. A Comparison of Statistical Significance Tests for Information Retrieval Evaluation. In *Proc. 16th International Conference on Information and Knowledge Management (CIKM 2007)*. ACM Press, New York, USA, 623–632.

[37] M. D. Smucker, J. Allan, and B. A. Carterette. 2009. Agreement Among Statistical Significance Tests forInformation Retrieval Evaluation at Varying Sample Sizes. In *Proc. 32nd Annual International ACM SIGIR (SIGIR 2009)*. ACM Press, New York, USA, 630–631.

[38] K. Spärck Jones. 1974. Automatic indexing. *Journal of Documentation* 30, 4 (1974), 393–432.

[39] T.D. Sterling. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *Journal of the American Statistical Association* 54, 285 (1959), 30–34.

[40] D. Szymkiewicz. 1934. Une contribution statistique à la géographie floristique. *Acta Societas Botanicorum Poloniae* 11, 3 (1934), 249–265.

[41] J. Tague, M. Nelson, and H. Wu. 1980. Problems in the Simulation of Bibliographic Retrieval Systems. In *Proceedings of the 3rd Annual ACM (SIGIR '80)*. Butterworth & Co., GBR, 236–255.

[42] J. M. Tague-Sutcliffe and J. Blustein. 1994. A Statistical Analysis of the TREC-3 Data. In *The Third Text REtrieval Conference (TREC-3)*. National Institute of Standards and Technology (NIST), Special Publication 500-225, Washington, USA, 385–398.

[43] J. W. Tukey. 1949. Comparing Individual Means in the Analysis of Variance. *Biometrics* 5, 2 (June 1949), 99–114.

[44] J. Urbano, H. Lima, and A. Hanjalic. 2019. Statistical Significance Testing in Information Retrieval: An Empirical Analysis of Type I, Type II and Type III Errors, See [28], 505–514.

[45] J. Urbano, M. Marrero, and D. Martín. 2013. A Comparison of the Optimality of Statistical Significance Tests for Information Retrieval Evaluation, See [21], 925–928.

[46] Julián Urbano and Thomas Nagler. 2018. Stochastic simulation of test collections: Evaluation scores. In *The 41st international ACM SIGIR conference on research & development in information retrieval*. 695–704.

[47] C. J. van Rijsbergen. 1979. *Information Retrieval* (2nd ed.). Butterworths, London, England.

[48] E. M. Voorhees. 2004. Overview of the TREC 2004 Robust Track. In *The Thirteenth Text REtrieval Conference Proceedings (TREC 2004)*. National Institute of Standards and Technology (NIST), Special Publication 500-261, Washington, USA.

[49] E. M. Voorhees and C. Buckley. 2002. The Effect of Topic Set Size on Retrieval Experiment Error. In *Proc. 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2002)*. ACM Press, New York, USA, 316–323.

[50] E. M. Voorhees, D. Samarov, and I. Soboroff. 2017. Using Replicates in Information Retrieval Evaluation. *ACM Transactions on Information Systems (TOIS)* 36, 2 (September 2017), 12:1–12:21.

[51] W. J. Wilbur. 1994. Non-parametric significance tests of retrieval performance comparisons. *Journal of Information Science* 20, 4 (August 1994), 270–284.

[52] J. Zobel. 1998. How Reliable are the Results of Large-Scale Information Retrieval Experiments. In *Proc. 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 1998)*. ACM Press, New York, USA, 307–314.