How to Measure the Reproducibility of System-oriented IR Experiments

Breuer, Timo; Ferro, Nicola; Fuhr, Norbert; Maistro, Maria; Sakai, Tetsuya; Schaer, Philipp; Soboroff, Ian

Published in:
SIGIR '20

DOI:
10.1145/3397271.3401036

Publication date:
2020

Document version
Peer reviewed version

Document license:
CC BY

Citation for published version (APA):
Breuer, T., Ferro, N., Fuhr, N., Maistro, M., Sakai, T., Schaer, P., & Soboroff, I. (2020). How to Measure the Reproducibility of System-oriented IR Experiments. In SIGIR '20: Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (pp. 349-358). Association for Computing Machinery. https://doi.org/10.1145/3397271.3401036
# How to Measure the Reproducibility of System-oriented IR Experiments

Timo Breuer  
TH Köln, Germany  
timo.breuer@th-koeln.de  

Nicola Ferro  
University of Padua, Italy  
ferro@dei.unipd.it  

Norbert Fuhr  
Universität Duisburg-Essen, Germany  
norbert.fuhr@uni-due.de  

Maria Maistro  
University of Copenhagen, Denmark  
mm@di.ku.dk  

Tetsuya Sakai  
Waseda University, Japan  
tetsuyasakai@acm.org  

Philipp Schaer  
TH Köln, Germany  
philipp.schaer@th-koeln.de  

Ian Soboroff  
NIST, USA  
ian.soboroff@nist.gov  

## ABSTRACT

Reproducibility and reproducibility of experimental results are primary concerns in all the areas of science and Information Retrieval (IR) is not an exception. Besides the problem of moving the field towards more reproducible experimental practices and protocols, we also face a severe methodological issue: we do not have any means to assess whether an experimental result is reproduced. Moreover, we lack any reproducibility-oriented dataset, which would allow us to validate our measures and which can also be used to develop future measures.

To address these issues, we compare several measures to objectively quantify to what extent an experiment has replicated or reproduced a system-oriented IR experiment. These measures operate at different levels of granularity, from the fine-grained comparison of ranked lists, to the more general comparison of the obtained effects and significant differences. Moreover, we also develop a reproducibility-oriented dataset, which allows us to validate our measures and which can also be used to develop future measures.

## 1 INTRODUCTION

We are today facing the so-called *reproducibility crisis* [3, 30] across all areas of science, where researchers fail to reproduce and confirm previous experimental findings. This crisis obviously involves also the more recent computational and data-intensive sciences [19, 29], including hot areas such as artificial intelligence and machine learning [22]. For example, Baker [3] reports that roughly 70% of researchers in physics and engineering fail to reproduce someone else’s experiments and roughly 50% fail to reproduce even their own experiments.

Information Retrieval (IR) is not an exception and researchers are paying more and more attention to what the reproducibility crisis may mean for the field, even more with the raise of the new deep learning and neural approaches [8, 10].

In addition to all the well-known barriers to reproducibility [19], a fundamental methodological question remains open: *When we say that an experiment is reproduced, what exactly do we mean by it?* The current attitude is some sort of “close enough”: researchers put any reasonable effort to understand how an approach was implemented and how an experiment was conducted and, after some (several) iterations, when they obtain performance scores which somehow resemble the original ones, they decide that an experimental result is reproduced. Unfortunately, IR completely lacks any means to objectively measure when reproduced is reproduced and this severely hampers the possibility both to assess to what extent an experimental result has been reproduced and to sensibly compare among different alternatives for reproducing an experiment.

This severe methodological impediment is not limited to IR but it has been recently brought up as a research challenge also in the 2019 report on “Reproducibility and Replicability in Science” by the US National Academies of Sciences, Engineering, and Medicine [29, p. 62]: “The National Science Foundation should consider investing in research that explores the limits of computational reproducibility in instances in which bitwise reproducibility is not reasonable in order to ensure that the meaning of consistent computational..."
results remains in step with the development of new computational hardware, tools, and methods. Another severe issue is that we lack any experimental collection specifically focused on reproducibility and this prevents us from developing and comparing measures to assess the extent of achieved reproducibility.

In this paper, we tackle both these issues. Firstly, we consider different measures which allow for comparing experimental results at different levels from most specific to most general: the ranked lists of retrieved documents; the actual scores of effectiveness measures; the observed effects and significant differences. As you can note these measures progressively depart more and more from the "bitwise reproducibility" [29] which in the IR case would mean producing exactly identical ranked lists of retrieved documents. Secondly, starting from TREC data, we develop a reproducibility-oriented dataset and we use it to compare the presented measures.

The paper is organized as follows: Section 2 discusses related work; Section 3 introduces the evaluation measures under investigation; Section 4 describes how we created the reproducibility-oriented dataset; Section 5 presents the experimental comparison of the evaluation measures; finally, Section 6 draws some conclusions and outlooks future work.

2 RELATED WORK
In defining what repeatability, replicability, reproducibility, and other of the so-called r-words are [31], De Roure [11] lists 21 r-words grouped in 6 categories, which range from scientific method to understanding and curation. In this paper, we broadly align with the definition of replicability and reproducibility currently adopted by the Association for Computing Machinery (ACM):2

- **Replicability (different team, same experimental setup):** the measurement can be obtained with stated precision by a different team using the same measurement procedure, the same measuring system, under the same operating conditions, in the same or a different location on multiple trials. For computational experiments, an independent group can obtain the same result using the author’s own artifacts;

- **Reproducibility (different team, different experimental setup):** the measurement can be obtained with stated precision by a different team, a different measuring system, in a different location on multiple trials. For computational experiments, an independent group can obtain the same result using artifacts which they develop completely independently.

Reproducibility Efforts in IR

There have been and there are several initiatives related to reproducibility in IR. Since 2015, the ECIR conference hosts a track dedicated to papers which reproduce existing studies, and all the major IR conferences ask an assessment of the ease of reproducibility of a paper in their review forms. The SIGIR group has started a task force [17] to define what reproducibility is in system-oriented and user-oriented IR and how to implement the ACM badging policy in this context. Fuhr [20, 21] urged the community to not forget about reproducibility and discussed reproducibility and validity in the context of the CLEF evaluation campaign. The recent ACM JDIQ special issue on reproducibility in IR [15, 16] provides an updated account of the state-of-the-art in reproducibility research as far as evaluation campaigns, collections, tools, infrastructures and analyses are concerned. The SIGIR 2015 RIGOR workshop [2] investigated reproducibility, inexplicable, and generalizability of results and held a reproducibility challenge for open source software [27]. The SIGIR 2019 OSIRRC workshop [7] conducted a replicability challenge based on Docker containers.

CENTRE3 is an effort across CLEF [14, 18], TREC [37], and NT-CIR [35] to run a joint evaluation activity on reproducibility. One of the goals of CENTRE was to define measures to quantify to which extent experimental results were reproduced. However, the low participation in CENTRE prevented the development of an actual reproducibility-oriented dataset and hampered the possibility of developing and validating measures for reproducibility.

Measuring Reproducibility

To measure reproducibility, CENTRE exploited: Kendall’s τ [25], to measure how close are the original and replicated list of documents; Root Mean Square Error (RMSE) [26], to quantify how close are the effectiveness scores of the original and replicated runs; and the Effect Ratio (ER) [35], to quantify how close are the effects of the original and replicated/reproduced systems.

We compare against and improve with respect to previous work within CENTRE. Indeed, Kendall’s τ cannot deal with rankings that do not contain the same elements; CENTRE overcomes this issue by considering the union of the original and replicated rankings and comparing with respect to it; we show how this is a somehow pessimistic approach, penalizing the systems and propose to use Rank-Biased Overlap (RBO) [40], since it is natively capable to deal with rankings containing different elements. Furthermore, we complement Effect Ratio (ER) with the new Delta Relative Improvement (DeltaRI) score, to better grasp replicability and reproducibility in terms of absolute scores and to provide a visual interpretation of the effects. Finally, we propose to test replicability and reproducibility with paired and unpaired t-test [38] respectively, and to use p-values as an estimate of replicability and reproducibility success.

To the best of our knowledge, inspired by the somehow unsuccessful experience of CENTRE, we are the first to systematically investigate measures for guiding replicability and reproducibility in IR, backing this with the development of a reproducibility-oriented dataset. As previously observed, there is a compelling need for reproducibility measures for computational and data-intensive sciences [29], being the largest body of knowledge focused on traditional lab experiments and metrology [24, 28], and we try here to start addressing that need in the case of IR.

3 PROPOSED MEASURES

We first introduce our notation. In all cases we assume that the original run r is available. For replicability (§ 3.1), both the original run r and the replicated run r’ contain documents from the original collection C. For reproducibility (§ 3.2), r denotes the original run on the original collection C, while r’ denotes the reproduced run on the new collection D. Topics are denoted by j ∈ {1, ..., n_doc} in C and j ∈ {1, ..., n_doc'} in D, while rank positions are denoted by i. M is any IR evaluation measure e.g., P@10, AP, nDCG, where the

---

1https://www.acm.org/publications/policies/artifact-review-badging, April 2018.
2https://www.centre-eval.org/
superscript $C$ or $D$ refers to the collection. $M^C_j(r)$ is the vector of length $n_C$ where each component, $M^C_j(r)_j$, is the score of the run $r$ with respect to the measure $M$ and topic $j$. $M^C(r)$ is the average score computed across topics.

### 3.1 Replicability

We evaluate replicability at different levels: (i) we consider the actual ordering of documents by using Kendall’s $\tau$ and Rank-Biased Overlap (RBO) [40]; (ii) we compare the runs in terms of effects with RMSE; (iii) we consider whether the overall effect can be replicated with Effect Ratio (ER) and Delta Relative Improvement (DeltaRI); and (iv) we compute statistical comparisons and consider the $p$-value of a paired t-test. While Kendall’s $\tau$, RMSE and ER were originally proposed for CENTRE, the other approaches has never been used for replicability.

It is worth mentioning that these approaches are presented from the most specific to the most general. Kendall’s $\tau$ and RBO compares the runs at document level, RMSE accounts for the performance at topic level, ER and DeltaRI focus on the overall performance by considering the average across topics, while the t-test can just inform us on the significant differences between the original and replicated runs. Moreover, perfect equality for Kendall’s $\tau$ and RBO implies perfect equality for RMSE, ER/DeltaRI and t-test, and perfect equality for RMSE implies perfect equality for ER/DeltaRI and t-test, while viceversa is in general not true.

As reference point, we consider the average score across topics of the original and replicated runs, called Average Retrieval Performance (ARP). Its delta represents the current “naïve” approach to replicability, simply contrasting the average scores of the original and replicated runs.

**Ordering of Documents.** Kendall’s $\tau$ is computed as follows [25]:

$$\tau_j(r, r') = \frac{P - Q}{\sqrt{(P + Q + U)(P + Q + V)}}$$

$$\bar{\tau}(r, r') = \frac{1}{n_C} \sum_{j=1}^{n_C} \tau_j(r, r')$$

where $\tau_j(r, r')$ is Kendall’s $\tau$ for the $j$-th topic, $P$ is the total number of concordant pairs (document pairs that are ranked in the same order in both vectors), $Q$ the total number of discordant pairs (document pairs that are ranked in opposite order in the two vectors), $U$ and $V$ are the number of ties, in $r$ and $r'$ respectively.

This definition of Kendall’s $\tau$ is originally proposed for permutations of the same set of items, therefore it is not directly applicable whenever two rankings do not contain the same set of documents. However, this is not the case of real runs, which often return different sets of documents. Therefore, as done in CENTRE@CLEF [14, 18], we consider the correlation with respect to the union of the rankings. We refer to this method as Kendall’s $\tau$ Union. The underlying idea is to compare the relative orders of documents in the original and replicated rankings. For each topic, we consider the union of $r$ and $r'$, by removing duplicate entries. Then we consider the rank positions of documents from the union in $r$ and $r'$, obtaining two lists of rank positions. Finally, we compute the correlation between these two lists of rank positions. Note that, whenever two rankings contain the same set of documents, Kendall’s $\tau$ in Eq. (1) and Kendall’s $\tau$ Union are equivalent. To better understand how Kendall’s $\tau$ Union is defined, consider two rankings: $r = [d_1, d_2, d_3]$ and $r' = [d_1', d_2, d_3]$, the union of $r$ and $r'$ is $[d_1, d_2, d_3, d_3']$, then the two lists of rank positions are $[1, 2, 3]$ and $[1, 2, 4]$ and the final Kendall’s $\tau$ is equal to 1. Similarly consider $r = [d_1, d_2, d_3, d_4]$ and $r' = [d_1, d_3, d_4, d_5]$, the union of $r$ and $r'$ is $[d_1, d_2, d_3, d_4, d_5]$, then the two lists of rank positions are $[1, 2, 3, 4]$ and $[2, 5, 3, 6]$ and the final Kendall’s $\tau$ is equal to 2/3.

We also consider Kendall’s $\tau$ on the intersection of the rankings instead of the union. As reported in [36], Kendall’s $\tau$ can be very noisy with small rankings and should be considered together with the size of the overlap between the 2 rankings. However, this approach does not inform us on the rank positions of the common documents. Therefore, to seamlessly deal with rankings possibly containing different documents and to account for their rank positions, we propose to use Rank-Biased Overlap (RBO) [40], which assumes $r$ and $r'$ to be infinite runs:

$$RBO_j(r, r') = (1 - \phi) \sum_{i=1}^{n} \phi^{i-1} \cdot A_i$$

$$RBO(r, r') = \frac{1}{n_C} \sum_{j=1}^{n_C} RBO_j(r, r')$$

where $RBO_j(r, r')$ is RBO for the $j$-th topic; $\phi \in [0, 1]$ is a parameter to adjust the measure top-heaviness: the smaller $\phi$, the more top-weighted the measure; and $A_i$ is the proportion of overlap up to rank $i$, which is defined as the cardinality of the intersection between $r$ and $r'$ up to $i$ divided by $i$. Therefore, RBO accounts for the overlap of two rankings and discounts the overlap while moving towards the end of the ranking, since it is more likely for two rankings to have a greater overlap when many rank positions are considered.

**Effectiveness.** As reported in CENTRE@NTCIR [35], we exploit Root Mean Square Error (RMSE) [26] to measure how close the effectiveness scores of the replicated and original runs are:

$$RMSE \left(M^C(r), M^C(r')\right) = \sqrt{\frac{1}{n_C} \sum_{j=1}^{n_C} (M^C_j(r) - M^C_j(r'))^2}$$

RMSE depends just on the evaluation measure and on the relevance label of each document, not on the actual documents retrieved by each run. Therefore, if two runs $r$ and $r'$ retrieve different documents, but with the same relevance labels, then RMSE is not affected and returns a perfect replicability score equal to 0; on the other hand, Kendall’s $\tau$ and RBO will be able to detect such differences.

Although RMSE and the naive comparison of ARP scores can be thought as similar approaches, by taking the squares of the absolute differences, RMSE penalizes large errors more. This can lead to different results, as shown in Section 5.

**Overall Effect.** In this case, we define a replication task from a different perspective, as proposed in CENTRE@NTCIR [35]. Given a pair of runs, $a$ and $b$, such that the advanced $a$-run has been reported to outperform the baseline $b$-run on the collection $C$, can another research group replicate the improvement of the advanced run over the baseline run on $C$? With this perspective, the per-topic
improvements in the original and replicated experiments are:

$$\Delta M_j^C = M_j^C(a) - M_j^C(b), \quad \Delta' M_j^C = M_j^C(a') - M_j^C(b')$$ (4)

where \(a'\) and \(b'\) are the replicated advanced and baseline runs respectively. Note that even if the \(a\)-run outperforms the \(b\)-run on average, the opposite may be true for some topics: that is, per-topic improvements may be negative.

Since IR experiments are usually based on comparing mean effectiveness scores, Effect Ratio (ER) [35] focuses on the replicability of the overall effect as follows:

$$ER(\Delta M^C, \Delta M^C) = \frac{\frac{1}{n} \sum \Delta M_j^C}{\frac{1}{n} \sum \Delta M_j^C}$$ (5)

where the denominator of ER is the mean improvement in the original experiment, while the numerator is the mean improvement in the replicated experiment. Assuming that the standard deviation for the difference in terms of measure \(M\) is common across experiments, ER is equivalent to the ratio of effect sizes (or standardised mean differences for the paired data case) [34]; hence the name.

ER \(\leq 0\) means that the replicated \(a\)-run failed to outperform the replicated \(b\)-run: the replication is a complete failure. If \(0 < ER < 1\), the replication is somewhat successful, but the effect is smaller compared to the original experiment. If \(ER = 1\), the replication is perfect in the sense that the original effect has been recovered as is. If \(ER > 1\), the replication is successful, and the effect is actually larger compared to the original experiment.

Note that having the same mean delta scores, i.e. \(ER = 1\), does not imply that the per-topic replication is perfect. For example, consider two topics \(i\) and \(j\) and assume that the original delta scores are \(\Delta M_j^C = 0.2\) and \(\Delta M_j^C = 0.8\) while the replicated delta scores are \(\Delta' M_j^C = 0.8\) and \(\Delta' M_j^C = 0.2\). Then ER for this experiment is equal to 1. While this difference is captured by RMSE or Kendall’s \(\tau\), which focus on a per-topic level, ER considers instead whether the sample effect size (standardised mean difference) can be replicated or not.

ER focuses on the effect of the \(a\)-run over the \(b\)-run, isolating it from other factors, but we may also want to account for absolute scores that are similar to the original experiment. Therefore, we propose to complement ER with Delta Relative Improvement (DeltaRI) and to plot ER against DeltaRI to visually interpret the replicability of the effects. We define DeltaRI as follows:

$$RI = \frac{M_j^C(a) - M_j^C(b)}{M_j^C(b)}, \quad RI' = \frac{M_j^C(a') - M_j^C(b')}{M_j^C(b')}$$ (6)

where \(RI\) and \(RI'\) are the relative improvements for the original and replicated runs and \(M_j^C(\cdot)\) is the average score across topics. Now let DeltaRI be \(ARI(RI, RI') = RI - RI'\). DeltaRI ranges in \([-1, 1]\), \(ARI = 0\) means that the relative improvements are the same for the original and replicated runs; when \(ARI > 0\), the replicated relative improvement is smaller than the original relative improvement, and in case \(ARI < 0\), it is larger. DeltaRI can be used in combination with ER, by plotting ER (x-axis) against DeltaRI (y-axis), as done in Figure 2. If \(ER = 1\) and \(ARI = 0\) both the effect and the relative improvements are replicated, therefore the closer a point to \((1, 0)\) the more successful the replication experiment. We can now divide the ER-DeltaRI plane in 4 regions, corresponding to the 4 quadrants of the cartesian plane:

- Region both 1: \(ER > 0\) and DeltaRI > 0, the replication is somehow successful in terms of effect sizes, but not in terms of absolute scores;
- Region 2: \(ER < 0\) and DeltaRI > 0, the replication is a failure both in terms of effect sizes and absolute scores;
- Region 3: both \(ER < 0\) and DeltaRI < 0, the replication is a failure in terms of effect sizes, but not in terms of absolute scores;
- Region 4: \(ER > 0\) and DeltaRI < 0, this means that the replication is somehow successful both in terms of effect sizes and absolute scores.

Therefore, the preferred region is Region 4, with the condition that the best replicability runs are close to \((1, 0)\).

**Statistical Comparison.** We propose to compare the original and replicated runs in terms of their statistical difference: we run a two-tailed paired t-test between the scores of \(r\) and \(r'\) for each topic in \(C\) with respect to an evaluation measure \(M\). The p-value returned by the t-test informs on the extent to which \(r\) is successfully replicated: the smaller the \(p\) value, the stronger the evidence that \(r\) and \(r'\) are significantly different, thus \(r'\) failed in replicating \(r\).

Note that the \(p\)-value does not inform on the overall effect, i.e. we may know that \(r'\) failed to replicate \(r\), but we cannot infer whether \(r'\) performed better or worse than \(r\).

### 3.2 Reproducibility

Differently from replicability, for reproducibility the original and reproduced runs are not obtained on the same collection (different documents and/or topic sets), thus the original run cannot be used for direct comparison with the reproduced run. As a consequence, Kendall’s \(\tau\), RBO, and RMSE in Section 3.1 cannot be applied to the reproducibility task. Therefore, hereinafter we focus on: (i) reproducing the overall effect with ER; (ii) comparing the original and reproduced runs with statistical tests.

**Overall Effect.** CENTRE@NTCIR [35] defines ER for reproducibility as follows: given a pair of runs, \(a\)-run and \(b\)-run, where the \(a\)-run has been reported to outperform the \(b\)-run on a test collection \(C\), can another research group reproduce the improvement on a different test collection \(D\)? The original per-topic improvements are the same as in Eq. (4), while the reproduced per-topic improvements are defined as in Eq. (4) by replacing \(C\) with \(D\). Therefore, the resulting Effect Ratio (ER) [35] is defined as follows:

$$ER(\Delta M^D, \Delta M^C) = \frac{\frac{1}{n} \sum n_j \Delta M_j^D}{\Delta M_j^C} = \frac{1}{\frac{1}{n} \sum n_j \Delta M_j^D}$$ (7)

where \(n_j\) is the number of topics in \(D\). Assuming that the standard deviation of a measure \(M\) is common across experiments, the above version of ER is equivalent to the ratio of effect sizes (or standardised mean differences for the two-sample data case) [34]; it can then be interpreted in a way similar to the ER for replicability. Note that since we are considering the ratio of the mean improvements instead of the mean of the improvements ratio, Eq. (7) can be applied also when the number of topics in \(C\) and \(D\) is different.
Similarly to the replicability case, ER can be complemented with DeltaRI, whose definition is the same of Eq. (6), but RI’ is computed over the new collection $D$, instead of the original collection $C$. DeltaRI has the same interpretation as in the replicability case, i.e. to show if the improvement in terms of relative scores in the reproduced experiment are similar to the original experiment.

**Statistical Comparison.** With a t-test, we can also handle the case when the original and the reproduced experiments are based on different datasets. In this case, we need to perform a two-tailed unpaired t-test to account, for the different subjects used in the comparison.

The unpaired t-test assumes equal variance and this is likely to not happen when, e.g., you have two different sets of topics in the two datasets. However, the unpaired t-test is known to be robust to such violations and Sakai [33] has shown that Welch’s t-test, which assumes unequal variance, may be less reliable when the sample sizes differ substantially and the larger sample has a substantially larger variance.

### 4 DATASET

To evaluate the measures in Section 3, we need a reproducibility-oriented dataset and, to the best of our knowledge, this is the first attempt to construct such a dataset. The use case behind our dataset is that of a researcher who tries to replicate the methods described in a paper and who also tries to reproduce those results on a different collection; the researcher uses the presented measures as a guidance to select the best replicated/reproduced run and understand when reproduced is reproduced. Therefore, to cover both replicability and reproducibility, the dataset should contain both a baseline and an advanced run. Furthermore, the dataset should contain runs with different “quality levels”, roughly meant as being more or less “close” to the original run, to mimic the different attempts of a researcher to get closer and closer to the original run.

We reimplement WCrobust04 and WCrobust0405, two runs submitted by Grossman and Cormack [23] to the TREC 2017 Common Core track [1]. WCrobust04 and WCrobust0405 rank documents by routing using profiles [32]. In particular, Grossman and Cormack extract relevance feedback from a training corpus, train a logistic regression classifier with tfidf-features of relevant documents to a topic, and rank documents of a target corpus by their probability of being relevant to the same topic. The baseline run and the advanced run differ by the training data used for the classifier – one single corpus for WCrobust04, two corpora for WCrobust0405. We replicate runs using The New York Times Corpus, our target corpus; we reproduce runs using Washington Post Corpus. It is a requirement that all test collections, i.e., those used for training as well as the target collection, share at least some of the same topics.

Our replicated runs cover 50 topics, whereas the reproduced runs cover 25 topics. Full details on the implementation can be found in [5] and in the public repository [3], which also contains the full dataset, consisting of 200 runs.

To generate replicated and reproduced runs, we systematically change a set of parameters and derive 4 constellations consisting of 20 runs each, for a total of 80 runs (40 runs for replicability and 40 runs for reproducibility). We call them constellations because, by gradually changing the way in which training features are generated and the classifier is parameterized, we obtain sets of runs which are further and further away from the original run in a somehow controlled way and, in Section 5.1, we will exploit this regularity to validate the behaviour of our measures. The 4 constellations are:

- $rpl_wcr04$,
- $rpl_wcr04_tol$,
- $rpl_wcr04_d$.

An alternative to our approach could be to artificially alter one or more existing runs by swapping and/or changing retrieved documents or, even, to generate artificial runs fully from scratch. However, these artificial runs would have had no connection with the principled way in which a researcher actually proceeds when trying to reproduce an experiment and with her/his need to get orientation during this process. As a result, an artificially constructed dataset would lack any clear use case behind it.

### 5 EXPERIMENTAL EVALUATION

We evaluate our measures in two ways. Firstly, using the first 4 “regular” constellations described in Section 4, we check that our measures behave as expected in these known cases, roughly speaking we check that they tend to increase/decrease as expected. Secondly, using all the constellations described in Section 4, we check that our measures actually provide different viewpoints on replicability/reproducibility by conducting a correlation analysis. To this end, as usual, we compute Kendall’s $\tau$ correlation among the rankings.

1. [https://github.com/irgroup/sigir2020-measure-reproducibility](https://github.com/irgroup/sigir2020-measure-reproducibility)
of runs produced by each of our measures. Whenever the correlation between two measures is very high, we can report just one measure, since the other will likely represent redundant information [41]; furthermore, as suggested by Voorhees [39], we consider two measures equivalent if their correlation is greater than 0.9, and noticeably different if Kendall’s $\tau$ is below 0.8.

As effectiveness measures used with ARP, RMSE and ER, we select Average Precision (AP) and Normalized Discounted Cumulated Gain (nDCG) with cut-off 1000 and P@10. Even if P@10 might be redundant [41], we want to investigate whether it is easier to replicate/reproduce an experiment with a set-based measure. RBO is computed with $\phi = 0.8$. Even if Webber et al. [40] instantiate RBO with $\phi \geq 0.9$, we exploit a lower $\phi$. Inspired by the analysis for Rank-Biased Precision (RBP) in Ferrante et al. [12], we select a lower $\phi$ to consider a less top-heavy measure, since for replicability we do not want to replicate just the top rank positions.

5.1 Validation of Measures

Case Study: Replicability. Table 1 reports the retrieval performance for the baseline b-run WCrobust04 and the replicability measures: Kendall’s $\tau$, RBO, RMSE, and the $p$-values returned by the paired $t$-test. The corresponding table for WCrobust04OS reports similar results and is included in an online appendix. We report ER in Table 2 and plot ER against DeltaRI in Figure 2, additional ER-DeltaRI plots are included in the online appendix.

In Table 1, low values for Kendall’s $\tau$ and RBO highlights how hard it is to accurately replicate a run at ranking level. Replicability runs achieve higher RBO scores than Kendall’s $\tau$, showing that RBO is somehow less strict.

RMSE increases almost consistently when the difference between ARP scores of the original and replicated runs decreases. In general, RMSE values of P@10 are larger compared to those of AP and nDCG, due to P@10 having naturally higher variance (since it also considers a lower cut-off). For the constellation rpl$_{wcr04}$-tsf and rpl$_{wcr04}$-C, RMSE with P@10 increases, even if the difference between ARP scores decreases. As pointed out in Section 3.1, this is due to RMSE which penalizes large errors. On the other hand, RMSE decreases almost consistently as the cut-off value increases, as shown in Figure 1b. As expected, if we consider the whole ranking, the replicability runs retrieve more relevant documents and thus achieve better RMSE scores.

As a general observation, it is easier to replicate a run in terms of RMSE rather than Kendall’s $\tau$ or RBO. This is further corroborated by the correlation results in Table 4, which shows low correlation between RMSE and Kendall’s $\tau$. Therefore, even if the original and the replicated runs place documents with the same relevance labels in the same rank positions, those documents are not the same, as shown in Figure 1a, where Kendall’s $\tau$ is computed at different cut-offs. This does not affect the system performance, but it might affect the user experience, which can be completely different.

For the paired $t$-test, as the difference in ARP decreases, $p$-value increases, showing that the runs are more similar. This is further validated by high correlation results reported in Table 4 between ARP and $p$-values. Recall that the numerator of the $t$-value is basically computing the difference in ARP scores, thus explaining the consistency of these results.

For rpl$_{wcr04}$-tsf and rpl$_{wcr04}$-C, RMSE and $p$-values are not consistent: RMSE increases, thus the error increases, but $p$-values also increase, thus the runs are considered more similar. As aforementioned, this happens because RMSE penalizes large errors per topic, while the $t$-statistic is tightly related to ARP scores.

Table 2 (left) reports ER scores for replicability runs. WCrobust04 is the baseline b-run, while WCrobust04OS is the advanced a-run, both of them on TREC Common Core 2017. Recall that, for ER, the closer the score to 1, the more successful the replication. ER behaves as expected: when the quality of the replicated runs deteriorates, ER scores tend to move further from 1. As for RMSE, we can observe that the extent of success for the replication experiments depends on the effectiveness measure. Thus, the best practice is to consider multiple effectiveness measures.

Note that, for the constellations of runs rpl$_{wcr04}$-tsf and rpl$_{-}wcr04$-C, there is no agreement among the best replication experiment when different effectiveness measures are considered. This trend is similar to the one observed with RMSE, $p$-values and delta in ARP. For example, for ER with P@10, the best replicability runs are rpl$_{wcr04}$-tsf3 and rpl$_{wcr04}$OS-ts3 but ER scores are not stable, while for AP and nDCG, ER values tend to move further from 1, as we deteriorate the replicability runs. Again, this is due to the high variance of P@10.

Figure 2 illustrates ER scores against DeltaRI for 2 constellations in Table 2 and the other constellations are included in the online appendix.
Table 1: Replicability results for WCrobust04: ARP, rank correlations, RMSE, and $p$-values returned by the paired $t$-test.

| run               | ARP      | Correlation | RMSE     | $p$-value |
|-------------------|----------|-------------|----------|-----------|
| WCrobust04        | 0.6460   | 0.3711      | 0.6371   | 0.007     |
| rpl_wcr04_tf_1    | 0.6920   | 0.3646      | 0.6172   | 0.0117    |
| rpl_wcr04_tf_2    | 0.6990   | 0.3624      | 0.6177   | 0.0096    |
| rpl_wcr04_tf_3    | 0.6828   | 0.3420      | 0.6011   | 0.0076    |
| rpl_wcr04_tf_4    | 0.6680   | 0.3106      | 0.5711   | 0.0037    |
| rpl_wcr04_tf_5    | 0.6220   | 0.2806      | 0.5365   | 0.0064    |
| rpl_wcr04_df_1    | 0.6700   | 0.3569      | 0.6145   | 0.0078    |
| rpl_wcr04_df_2    | 0.6560   | 0.3425      | 0.5939   | 0.0073    |
| rpl_wcr04_df_3    | 0.6600   | 0.3049      | 0.5692   | 0.0072    |
| rpl_wcr04_df_4    | 0.5220   | 0.2519      | 0.5058   | 0.0048    |
| rpl_wcr04_df_5    | 0.4480   | 0.2121      | 0.4512   | 0.0019    |
| rpl_wcr04_tol_1   | 0.6700   | 0.3479      | 0.5992   | 0.0033    |
| rpl_wcr04_tol_2   | 0.5680   | 0.2877      | 0.5049   | 0.0061    |
| rpl_wcr04_tol_3   | 0.3700   | 0.1812      | 0.3269   | 0.0066    |
| rpl_wcr04_tol_4   | 0.2180   | 0.0903      | 0.1728   | 0.0066    |
| rpl_wcr04_tol_5   | 0.0700   | 0.0088      | 0.0379   | 0.0012    |
| rpl_wcr04_C_1     | 0.7020   | 0.3671      | 0.6191   | 0.0039    |
| rpl_wcr04_C_2     | 0.6960   | 0.3717      | 0.6244   | 0.0021    |
| rpl_wcr04_C_3     | 0.6840   | 0.3532      | 0.6093   | 0.0096    |
| rpl_wcr04_C_4     | 0.6240   | 0.3168      | 0.5761   | 0.0073    |
| rpl_wcr04_C_5     | 0.6140   | 0.3085      | 0.5689   | 0.0068    |

appendix. Recall that in Figure 2, the closer a point to the reference (1, 0), the better the replication experiment, both in terms of effect sizes and absolute differences.

The ER-DeltaR plot, can be used as a visual tool to guide researchers on the exploration of the “space of replicability” runs. For example, in Figure 2a, for AP and nDCG the point (1, 0) is reached from Region 4, which is somehow the preferred region, since it corresponds to successful replication both in terms of effect sizes and relative improvements. Conversely, in Figure 2b, it is clear that for AP the point (1, 0) is reached from Region 1, which corresponds to somehow a successful replication in terms of effect sizes, but not in terms of relative improvements.

Case Study: Reproducibility. For reproducibility, Table 3 reports ARP and $p$-values in terms of $P@10$, AP, and nDCG, for the runs reproducing WCrobust04 on TREC Common Core 2018. The corresponding table for WCrobust04 is included in the online appendix. Note that, in this case we do not have the original run scores, so we cannot directly compare ARP values. This represents the main challenge when evaluating reproducibility runs.

From $p$-values in Table 3, we can conclude that all the reproducibility runs are statistically significantly different from the original run, being the highest $p$-value just 0.005. Therefore, it seems that none of the runs successfully reproduced the original run.

However, this is likely due to the two collections being too different, which in turn makes the scores distribution also different. Consequently the $t$-test considers all the distributions as significantly different. To validate this hypothesis, we carried out an unpaired $t$-test between pairs of replicability and reproducibility runs in the 4 different constellations. This means that each pair of runs is generated by the same system on two different collections. The $p$-values for this experiment are reported only in the online appendix. Again, the majority of the runs are considered statistically different, except for a few cases for rpl_wcr04_df and rpl_wcr04_tol, which exhibit higher $p$-values also in Table 3. This shows that, depending on the collections, the unpaired $t$-test can fail in correctly detecting reproduced runs.

Table 2 (right) reports ER scores for replicability runs. At a first sight, we can see that ER scores are much lower than for the replicability case. If it is hard to perfectly replicate an experiment, it is even harder to perfectly reproduce it.

This is illustrated in the ER-DeltaR plot in Figure 3. In Figure 3a the majority of the points are far from the best reproduction (1, 0), even if they are in region 4. In Figure 3b just one point is in the preferred region 4, while many points are in region 2, that is failure both in reproducing the effect size and the relative improvement.

5.2 Correlation Analysis

Reproducibility. Note that for some measures, namely Kendall’s $\tau$, RBO, $p$-value, the higher the score the better the replicated run, conversely for RMSE and Delta ARP (absolute difference in ARP), the lower the score the better the replicated run. Thus, before computing the correlation among measures, we ensure that all the measure scores are consistent with respect to each other. Practically we consider the opposite of $\tau$, RBO and $p$-values, and for ER we consider $1 - ER$, since the closer its score to 1, the better the replicability performance.

Table 4 reports Kendall’s $\tau$ correlation for replicability measures on the set of runs replicating WCrobust04 (upper triangle, white
Table 2: ER results for replicability and reproducibility: the a-run is WCrobust0405 on TREC Common Core 2017; the b-run is WCrobust04, for replicability on TREC Common Core 2017, for reproducibility on TREC Common Core 2018.

| run         | replicability | reproductibility |
|-------------|---------------|------------------|
|             | P@10  AP  nDCG | P@10  AP  nDCG  |
| rpl_tf_1    | 0.8077 1.0330 1.1724 | 1.1923 1.2724 2.0299 |
| rpl_tf_2    | 0.7308 1.0347 1.1336 | 0.9615 1.3195 2.2139 |
| rpl_tf_3    | 0.9038 1.3303 1.3751 | 1.5000 1.5616 2.5365 |
| rpl_tf_4    | 0.6346 1.4719 1.5703 | 1.4231 1.9493 2.9317 |
| rpl_tf_5    | 1.1346 1.5955 1.8221 | 1.5385 1.7010 3.0569 |
| rpl_df_1    | 0.9615 0.9999 1.1006 | 0.4615 0.7033 0.9547 |
| rpl_df_2    | 1.0192 0.9207 1.0656 | 0.4231 0.4934 0.6586 |
| rpl_df_3    | 1.0385 0.8016 1.0137 | 0.1923 0.5429 1.0607 |
| rpl_df_4    | 0.9615 0.5911 0.8747 | 0.3846 0.5136 0.8333 |
| rpl_df_5    | 0.8654 0.3506 0.6459 | 0.3846 0.4857 0.7260 |
| rpl_tol_1   | 1.0769 1.2013 1.3455 | 0.5769 0.6574 0.8780 |
| rpl_tol_2   | 1.3269 1.4946 1.9290 | 0.8077 0.5194 0.8577 |
| rpl_tol_3   | 1.6854 2.1485 2.8496 | 2.0000 1.4524 2.9193 |
| rpl_tol_4   | 2.0962 2.2425 3.3213 | 2.3846 2.1242 3.9092 |
| rpl_tol_5   | 1.2500 1.0469 1.8504 | 0.2692 0.1116 0.5595 |
| rpl_C_1     | 0.6346 0.6300 0.8901 | 2.1538 1.8877 3.7777 |
| rpl_C_2     | 0.8077 0.7361 0.9240 | 2.2308 1.9644 3.8621 |
| rpl_C_3     | 0.8654 1.1195 1.2092 | 2.3846 2.2743 4.2783 |
| rpl_C_4     | 0.9231 1.1642 1.2911 | 0.6538 0.7316 1.0403 |
| rpl_C_5     | 0.8846 1.1214 1.2542 | 0.5769 0.6915 0.9741 |

Figure 3: Reproducibility: ER on the x-axis against DeltaRI on the y-axis.

(a) rpd_tf runs.  (b) rpd_df runs.

Table 3: Reproducibility: ARP and p-value (unpaired t-test), for WCrobust04. The original runs are on TREC Common Core 2017, and reproduced runs on TREC Common Core 2018.

| run         | ARP  | p-value  |
|-------------|------|----------|
|             | P@10 | AP  nDCG | P@10  AP  nDCG  |
| rpd_tf_1    | 0.3680 0.1619 0.3876 | 7.0E-04 6.9E-06 6.9E-06 |
| rpd_tf_2    | 0.3760 0.1628 0.3793 | 9.0E-04 8.6E-06 4.6E-06 |
| rpd_tf_3    | 0.3280 0.1468 0.3587 | 8.5E-05 1.6E-08 8.0E-07 |
| rpd_tf_4    | 0.3040 0.1180 0.3225 | 2.0E-05 3.0E-08 1.8E-08 |
| rpd_tf_5    | 0.2920 0.1027 0.2854 | 1.5E-05 6.0E-09 4.0E-10 |
| rpd_df_1    | 0.4240 0.1895 0.4543 | 0.005 8.0E-05 3.0E-04 |
| rpd_df_2    | 0.4290 0.1972 0.4727 | 0.003 1.4E-04 9.0E-04 |
| rpd_df_3    | 0.3880 0.1757 0.4304 | 0.001 2.0E-05 8.0E-05 |
| rpd_df_4    | 0.3360 0.1458 0.4000 | 7.0E-05 8.0E-07 6.0E-06 |
| rpd_df_5    | 0.2960 0.1140 0.3495 | 9.0E-06 1.0E-08 1.0E-07 |
| rpd_tol_1   | 0.4200 0.1872 0.4469 | 0.005 6.0E-05 2.0E-04 |
| rpd_tol_2   | 0.3960 0.1769 0.4134 | 0.002 3.0E-05 5.0E-05 |
| rpd_tol_3   | 0.2040 0.0987 0.2365 | 7.0E-08 8.0E-09 1.0E-10 |
| rpd_tol_4   | 0.0720 0.0183 0.0572 | 1.0E-12 5.0E-14 3.0E-22 |
| rpd_tol_5   | 0.0200 0.0007 0.0048 | 5.0E-16 1.0E-15 3.0E-27 |
| rpd_C_1     | 0.2600 0.1228 0.2786 | 5.0E-06 3.0E-07 2.0E-08 |
| rpd_C_2     | 0.2600 0.1216 0.2790 | 5.0E-06 2.0E-07 2.0E-08 |
| rpd_C_3     | 0.2360 0.0969 0.2507 | 8.0E-07 7.0E-09 5.0E-10 |
| rpd_C_4     | 0.3600 0.1609 0.4095 | 3.0E-04 4.0E-06 1.0E-05 |
| rpd_C_5     | 0.3520 0.1565 0.4026 | 2.0E-04 2.0E-06 8.0E-06 |

different to compare the overall average or the performance score topic by topic, as also shown by P@10 in Table 1. Furthermore, the correlation between RMSE instantiated with AP and nDCG is high, above 0.9, this is due to AP and nDCG being highly correlated, as also shown by the correlation between ARP with AP and nDCG (above 0.90) and between p-values with AP and nDCG (above 0.91).

When using the same performance measure, ARP and p-values approaches are highly correlated, even if from Table 1 several runs have small p-values and are statistically different. As mentioned in Section 5.1, the numerator of the t-stat is Delta ARP, and likely due to low variance, Delta ARP and p-values are tightly related. As explained in Section 3.1, ER takes a different perspective when evaluating replicability runs. This is corroborated by correlation results, which show that this measure has low correlation with ARP and any other evaluation approach. Indeed, replicating the overall improvement over a baseline, does not mean that there is perfect replication on each topic. Moreover, even the correlation among ER instantiated with different measures is low, which means that a mean improvement over the baseline in terms of AP does not necessarily correspond to a similar mean improvement for nDCG.

Reproducibility. For reproducibility we can not compare against ARP: since the original and reproduced runs are defined on different background and WCrobust0405 (lower triangle, turquoise background). The correlation between ARP and τ is low, below 0.29, and higher for RBO 0.70. This validates the findings from Section 5.1, showing that Kendall’s τ assumes a totally different perspective when evaluating replicability runs. Between τ and RBO, RBO correlates more with ARP than τ, especially with respect to AP and nDCG. Also, τ and RBO are low correlated with respect to each other. This is due to RBO being top-heavy, as AP and nDCG, while Kendall’s τ considers each rank position as equally important.

The correlation among ARP and RMSE is higher, especially when the same measure is considered by both ARP and RMSE. Nevertheless, the correlation is always lower than 0.86, showing that it is
We faced the core issue of investigating measures to determine to which gather and make available runs by their participants.

We found that replicability measures behave as expected and are robust to the different levels of granularity and the first reproducibility-oriented dataset. Due to the lack of a reproducibility-oriented dataset, these measures have never been validated so far.

We investigated measures that behave as expected and consistently; in particular, RMSE provides more meaningful comparisons than Kendall’s \( \tau \). RMSE properly indicates whether we obtained a similar level of performance; finally, both ER/DeltaRI and the paired t-test successfully determine whether the same effects are replicated. On the other hand, quantifying reproducibility is more challenging and, while ER/DeltaRI is still able to provide sensible insights, the unpaired t-test seems to be too sensitive to the differences among the experimental collections.

As a suggestion to improve our community practices, it is important to always provide not only the source code but also the actual run, as to enable precise checking for replicability; luckily, this is already happening when we operate within evaluation campaigns which gather and make available runs by their participants.

In future work, we will explore more advanced statistical methods to quantify reproducibility in a reliable way. Moreover, we will investigate how replicability and reproducibility are related to user experience. For example, a perfectly replicated run in terms of RMSE, but with low RBO, presents different documents to a user and this might greatly affect her/his experience. Therefore, we need to better understand which replicability/reproducibility level is needed to not impact (too much) on the user experience.

Acknowledgments. This paper is partially supported by AMAOS (Advanced Machine Learning for Automatic Omni-Channel Support), funded by Innovationsfonden, Denmark, and by DFG (German Research Foundation, project no. 407518790).
