How to Measure the Reproducibility of System-oriented IR Experiments

SIGIR '20, July 25–30, 2020, Virtual Event, China

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

Introduction

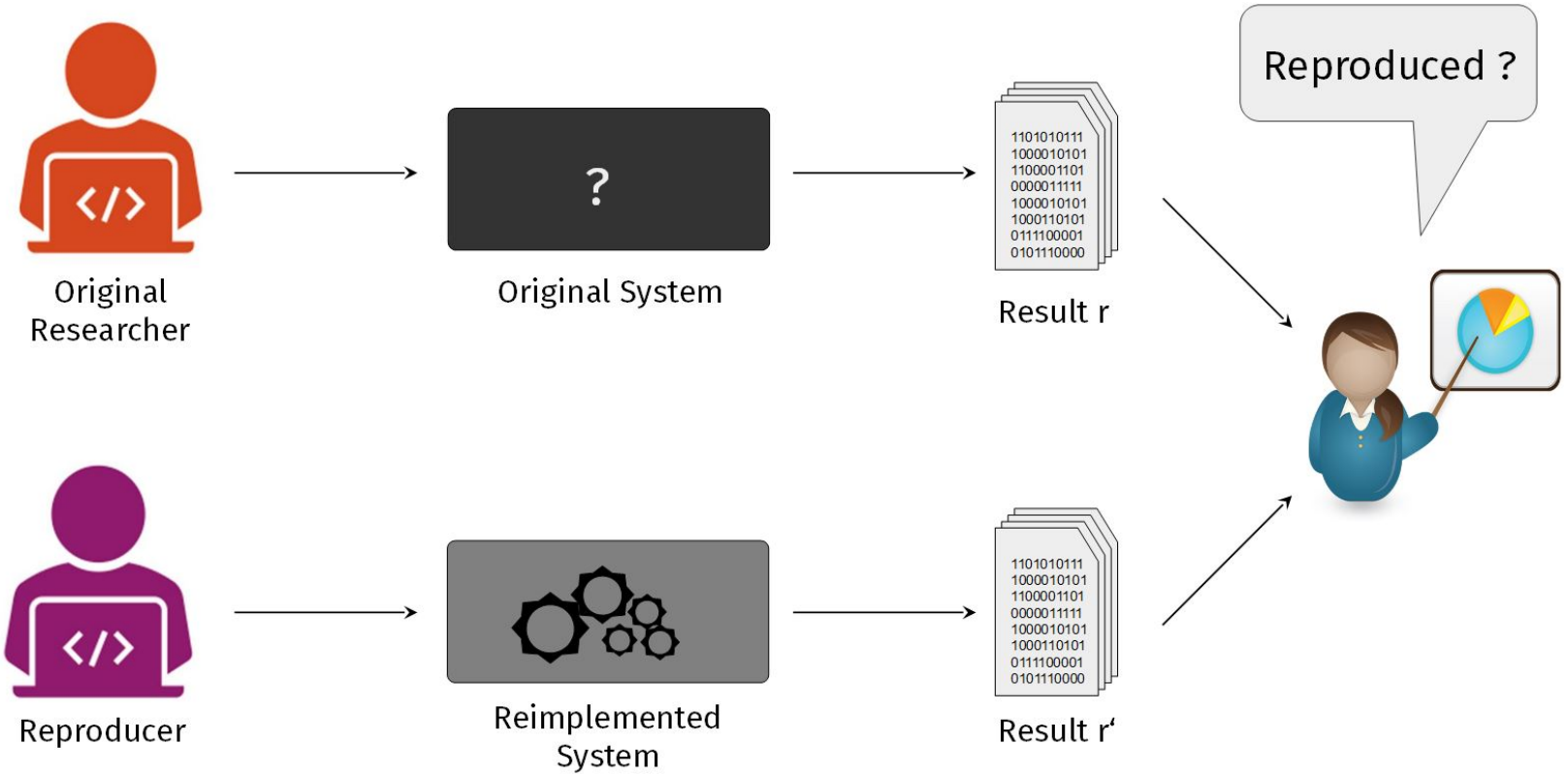
Why Reproducibility Matters

"An experimental result is not fully established unless it can be independently reproduced."[1]

"More than **70%** of researchers have tried and failed to replicate another scientist's experiments, and more than **half** have failed to repeat their own experiments."
M.Baker, *Nature* [1]

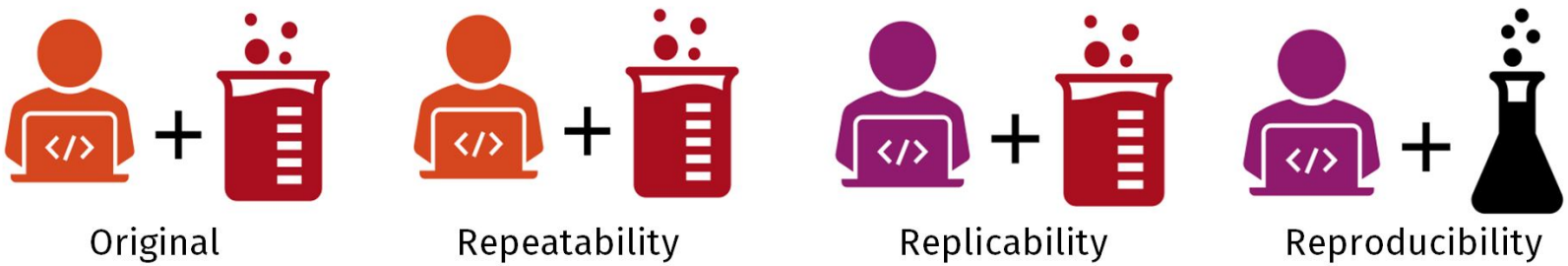
[1] `acm.org/publications/policies/artifact-review-badging`

Reproducibility Issues in Information Retrieval



Terminology

ACM *Artifact and Review Badging*[2]:



[2] acm.org/publications/policies/artifact-review-badging

Terminology

Application to system-oriented IR experiments[3] :

	System	Collection
Replicability	Reimplemented	Original
Reproducibility	Reimplemented	New

[3] centre-eval.org/

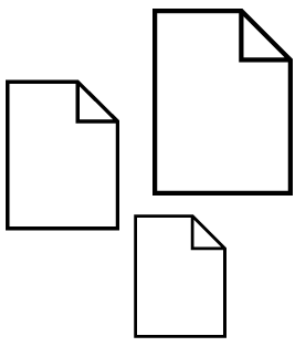
Goal

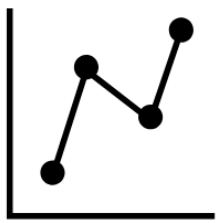
Investigate measures and methodologies for quantifying different levels of replication and reproduction.

Approach

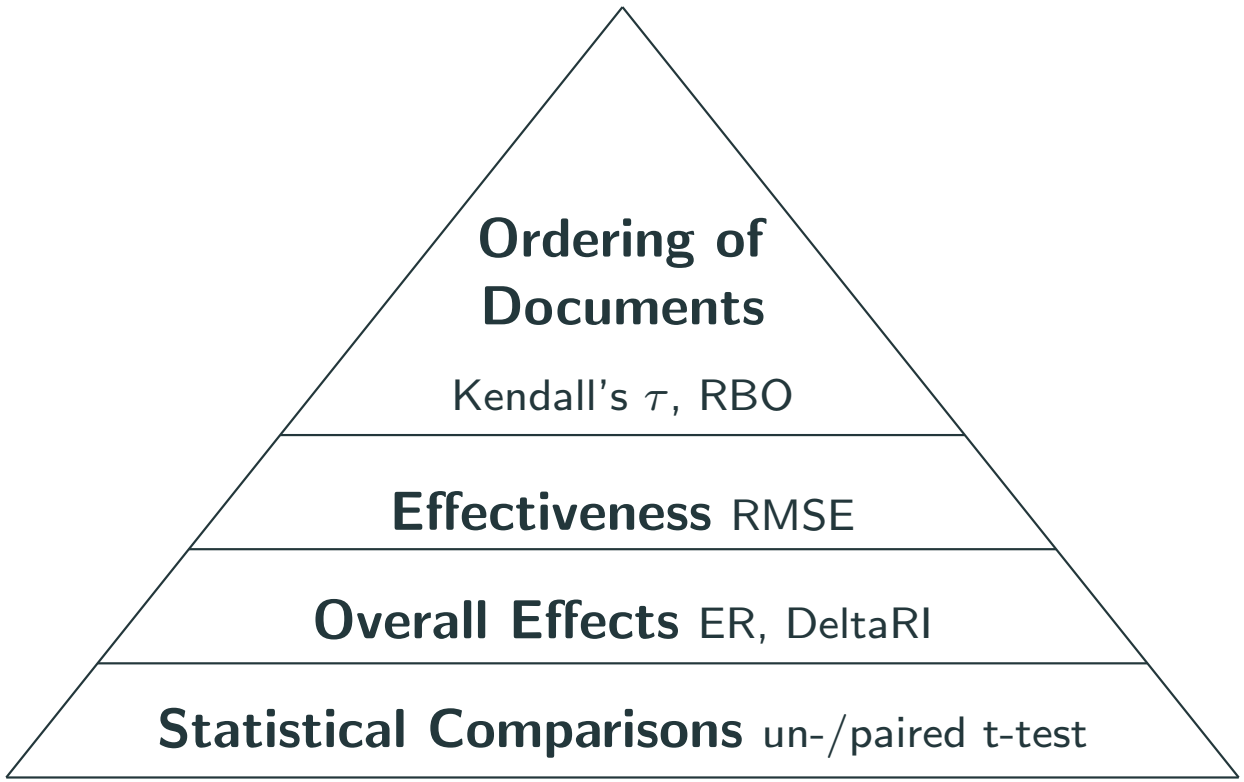
Outline

- Define a set of adequate **measures**
- Derive replica-/reproducibility-oriented **dataset**
- **Validate** measures with the dataset





Levels of Replication and Reproduction



Measures

Ordering of Documents

Kendall's τ [5]

- Compares permutations of the same set
- Kendall's τ Union [2, 3] lowers restriction by comparing "lists of ranks"

Rank-biased Overlap (RBO) [7]

- Compares list with possibly different documents
- Accounts for rank position

Effectiveness

CENTRE@CLEF [2, 3] exploits **Root Mean Square Error (RMSE)**:

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

M - Any IR evaluation measure (e.g. P@10, AP, nDCG)
$M^C(r)$ - Vector where each component is the score respect to the topic j

- RMSE is affected by the **relevance label**, not the actual document
- **Penalization of larger errors**

Overall Effects

Effect Ratio (ER) is introduced in CENTRE@NTCIR [6]:

$$\mathrm{ER}\left(\Delta' M^C, \Delta M^C\right) = \frac{\overline{\Delta' M^C}}{\overline{\Delta M^C}} = \frac{\frac{1}{n_C}\sum_{j=1}^{n_C} \Delta' M_j^C}{\frac{1}{n_C}\sum_{j=1}^{n_C} \Delta M_j^C}$$

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

Delta Relative Improvement (DeltaRI):

$$\Delta\mathrm{RI}(\mathrm{RI}, \mathrm{RI}') = \mathrm{RI} - \mathrm{RI}'$$

$$\mathrm{RI} = \frac{\overline{M^C(a)} - \overline{M^C(b)}}{\overline{M^C(b)}}, \qquad \mathrm{RI}' = \frac{\overline{M^C(a')} - \overline{M^C(b')}}{\overline{M^C(b')}}$$

a, a' - original and replicated/reproduced advanced run

b, b' - original and replicated/reproduced baseline run

Statistical Comparisons

Two-tailed t-tests

- t-test with r and r' for each topic j
- p-value gives evidence about significant differences
- no information about better or worse performance

Replicability

- same collection
- two-tailed **paired** t-test

Reproducibility

- new collection
- two-tailed **unpaired** t-test

Dataset

Reimplementation

- Reimplementations of `WCrobust04` (b-run) and `WCrobust0405` (a-run) by Grossman and Cormack [4]
- Automatic routing runs based on tfidf features and a logistic regression classifier
- Public repository[4]
- Replicability: *New York Times Corpus* (Core '17)
- Reproducibility: *TREC Washington Post Corpus* (Core '18)

[4] `github.com/irgroup/sigir2020-measure-reproducibility`

Dataset

The dataset[5] contains 200 different run constellations with varying parameters and processing steps. In our study we focus on four types of parameter variations:

- `rpl_wcr04_tf`
- `rpl_wcr04_df`
- `rpl_wcr04_tol`
- `rpl_wcr04_C`

[5] `zenodo.org/record/3856042`

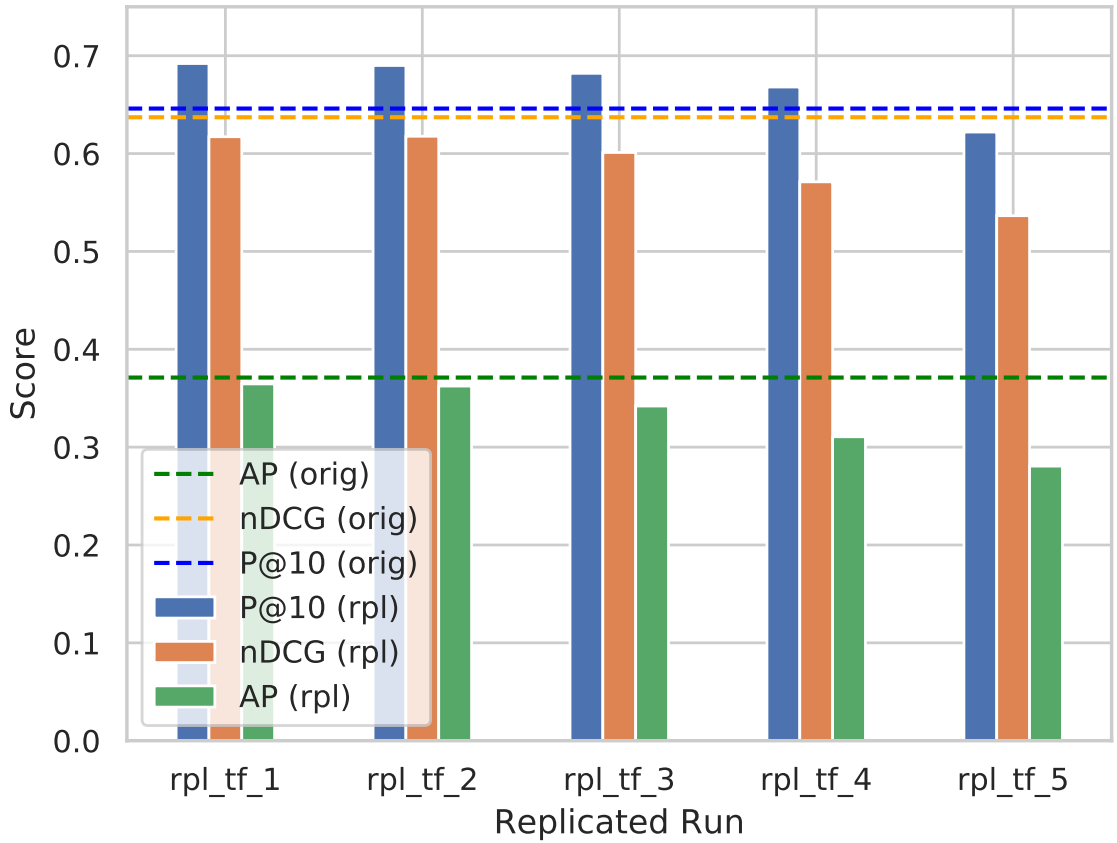
Experimental Evaluation

Experimental Evaluation

- **Part I**: Validation of Measures
- **Part II**: Correlation Analysis
- Included examples are based on replicated and reproduced `WCrobust04` and parameter variations with regards to `rpl_wcr04_tf`!

Replicability - Average Retrieval Performance (ARP)

Figure: Replicability results for the ARP of `WCrobust04`.

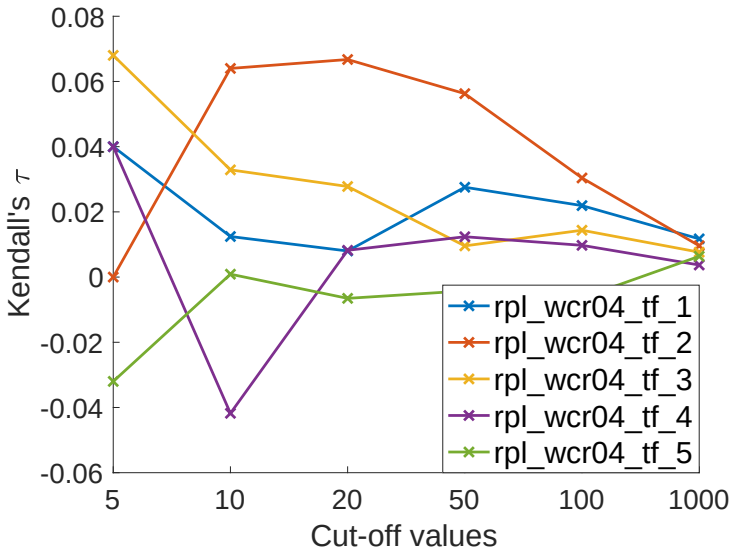


Replicability - Ordering of Documents

Table: Replicability results for the rank correlations of WCrobust04.

run	ARP nDCG	Correlation τ	Correlation RBO
WCrobust04	0.6371	1	1
rpl_wcr04_tf_1	0.6172	0.0117	0.5448
rpl_wcr04_tf_2	0.6177	0.0096	0.5090
rpl_wcr04_tf_3	0.6011	0.0076	0.4372
rpl_wcr04_tf_4	0.5711	0.0037	0.3626
rpl_wcr04_tf_5	0.5365	0.0064	0.2878

Figure: Kendall's τ of the replicated WCrobust04.

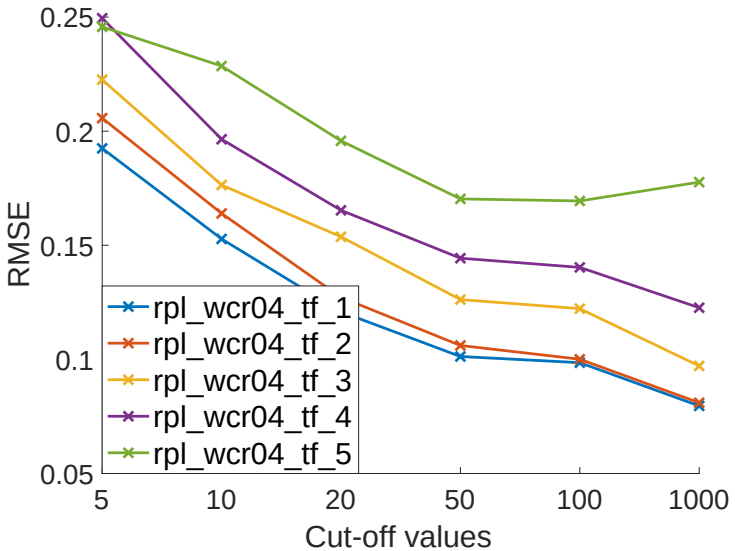


Replicability - Effectiveness

Table: Replicability results for the RMSE of `WCrobust04`.

run	ARP_{nDCG}	$RMSE_{nDCG}$
`WCrobust04`	0.6371	0
`rpl_wcr04_tf_1`	0.6172	0.0796
`rpl_wcr04_tf_2`	0.6177	0.0810
`rpl_wcr04_tf_3`	0.6011	0.0971
`rpl_wcr04_tf_4`	0.5711	0.1226
`rpl_wcr04_tf_5`	0.5365	0.1777

Figure: $RMSE_{nDCG}$ of the replicated `WCrobust04`.



Replicability - Overall Effects

Figure: ER results for replicability.

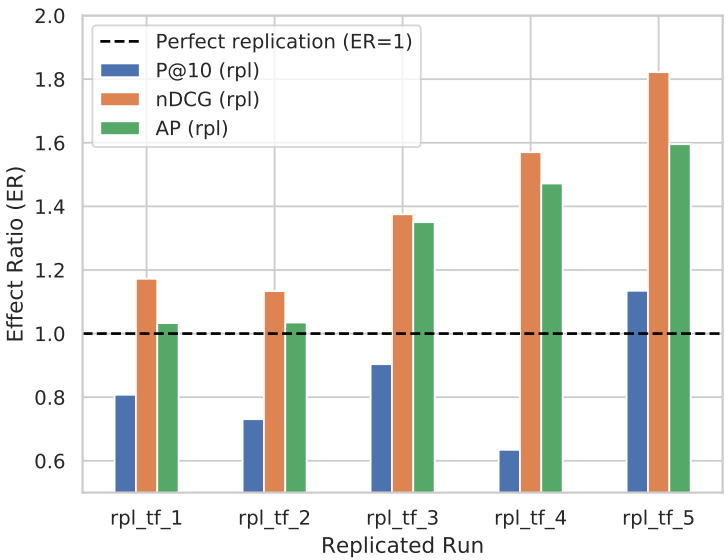
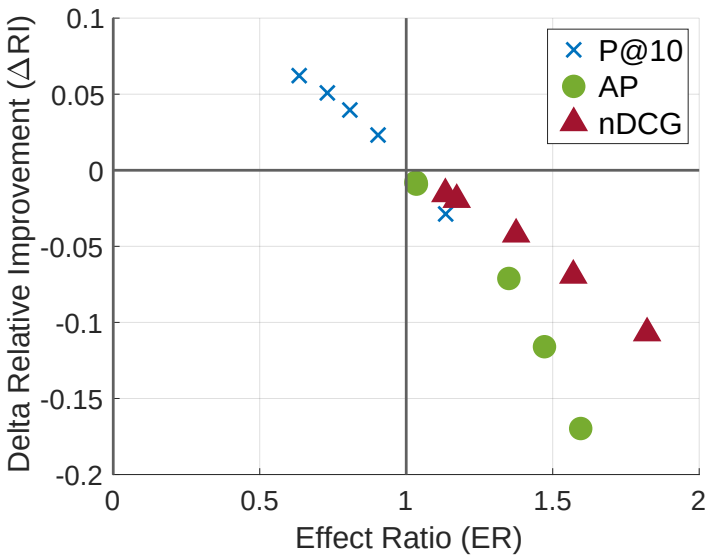


Figure: Replicated runs with varying `tf`-parameters.



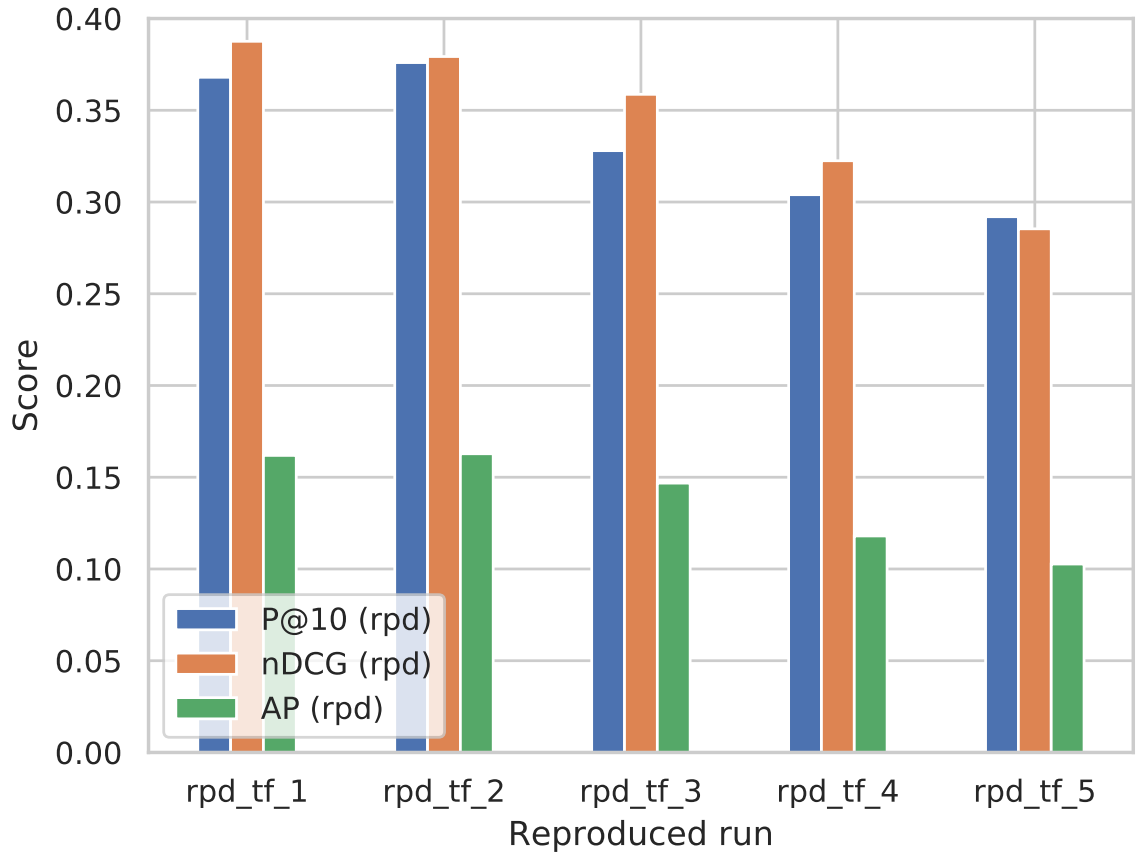
Replicability - Statistical Comparisons

Table: *p*-values returned by the two-tailed paired *t*-test in comparison to ARP and RMSE.

run	*nDCG* ARP	*nDCG* RMSE	*p*-value
WCrobust04	0.6371	0	1
rpl_wcr04_tf_1	0.6172	0.0796	0.077
rpl_wcr04_tf_2	0.6177	0.0810	0.090
rpl_wcr04_tf_3	0.6011	0.0971	0.007
rpl_wcr04_tf_4	0.5711	0.1226	$4E-05$
rpl_wcr04_tf_5	0.5365	0.1777	$1E-05$

Reproducibility - Average Retrieval Performance (ARP)

Figure: Reproducibility results for the ARP of `WCrobust04`.



Reproducibility - Overall Effects

Figure: ER results for reproducibility.

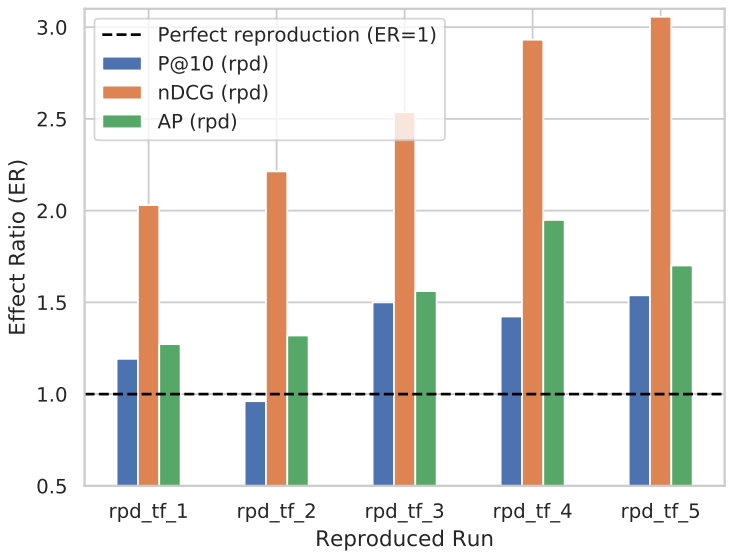
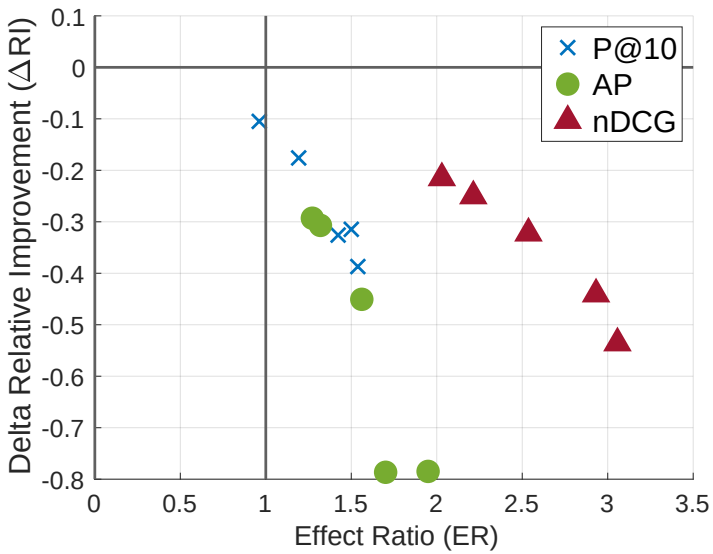


Figure: Reproduced runs with varying `tf`-parameters.



Reproducibility - Statistical Comparisons

Table: p-values returned by the two-tailed unpaired t-test.

run	*nDCG*	
	ARP	p-value
`rpd_wcr04_tf_1`	0.3876	$6E-06$
`rpd_wcr04_tf_2`	0.3793	$4E-06$
`rpd_wcr04_tf_3`	0.3587	$8E-07$
`rpd_wcr04_tf_4`	0.3225	$1E-08$
`rpd_wcr04_tf_5`	0.2854	$4E-10$

Correlation Analysis - Replicability

Table: Correlation among different measures for runs replicating `WCrobust04` (white background); and runs replicating `WCrobust0405` (turquoise background).

	Delta ARP			Correlation		RMSE			*p*-value			ER		
	P@10	AP	nDCG	τ	RBO	P@10	AP	nDCG	P@10	AP	nDCG	P@10	AP	nDCG
Δarp_P@10	-	0.4175	0.3979	0.2456	0.3684	0.3419	0.4552	0.4290	0.9156	0.3668	0.3700	0.2348	0.1752	0.0884
Δarp_AP	0.4535	-	0.9118	0.2718	0.7045	0.5209	0.8514	0.8090	0.3855	0.8841	0.8596	0.2145	0.3012	0.3731
Δarp_nDCG	0.4716	0.9363	-	0.2882	0.6555	0.5339	0.8580	0.8547	0.3463	0.8318	0.8302	0.2374	0.3208	0.4318
τ	0.2620	0.2865	0.2620	-	0.2180	0.2788	0.2702	0.2898	0.2434	0.2376	0.2457	0.1834	0.2718	0.2098
RBO	0.3946	0.6637	0.6457	0.3584	-	0.6026	0.7616	0.6898	0.3201	0.6376	0.6490	0.3307	0.2049	0.3029
RMSE_P@10	0.5420	0.6713	0.7089	0.3213	0.7433	-	0.6239	0.5944	0.2544	0.4080	0.4129	0.3452	0.2706	0.3753
RMSE_AP	0.5076	0.7747	0.8188	0.3224	0.7910	0.8136	-	0.8988	0.4034	0.7355	0.7273	0.2734	0.3453	0.4171
RMSE_nDCG	0.4666	0.7616	0.8188	0.3094	0.7682	0.8054	0.9184	-	0.3806	0.7127	0.6849	0.2767	0.3649	0.4498
p_value_P@10	0.8393	0.3694	0.3645	0.2566	0.2877	0.3790	0.3743	0.3400	-	0.3740	0.3593	0.2129	0.1486	0.0327
p_value_AP	0.3913	0.8498	0.7927	0.2506	0.5657	0.5470	0.6245	0.6180	0.3564	-	0.9135	0.1736	0.2343	0.2898
p_value_nDCG	0.3848	0.8416	0.7845	0.2424	0.5543	0.5356	0.6196	0.6033	0.3384	0.9069	-	0.2178	0.2163	0.3110
ER_P@10	0.0739	0.2652	0.2767	0.2227	0.3537	0.3108	0.3193	0.3144	0.0459	0.1817	0.1867	-	0.2833	0.1736
ER_AP	0.3013	0.2963	0.3078	0.1673	0.2343	0.3312	0.3551	0.3420	0.2599	0.1886	0.1706	0.2833	-	0.3992
ER_nDCG	0.2718	0.2767	0.3143	0.1216	0.2669	0.3377	0.3747	0.3551	0.1553	0.1494	0.1706	0.1736	0.3992	-

Correlation Analysis - Replicability

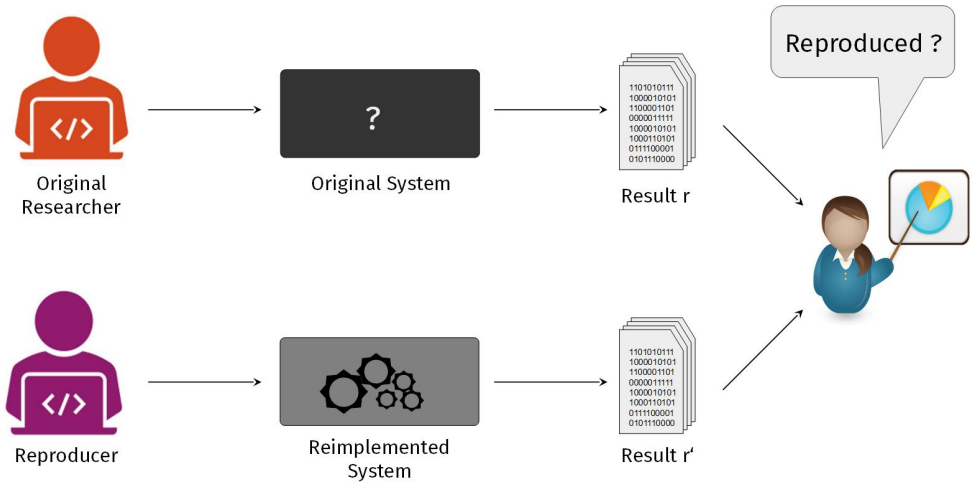
- RBO has a higher correlation with (top-heavy) ARP than Kendall's τ (with ARP)
- Correlation between RMSE and ΔARP is high
- Correlation between p-values and ΔARP is high
  (when using the same performance measure)
- ER has a low correlation with other measures

Correlation Analysis - Reproducibility

Table: Correlation among different measures for runs reproducing `WCrobust04` (white background); and runs reproducing `WCrobust0405` (turquoise background).

	p-value			ER		
	P@10	AP	nDCG	P@10	AP	nDCG
p_value_P@10	-	0.8545	0.8446	−0.2050	−0.1153	0.0025
p_value_AP	0.8168	-	0.8694	−0.1743	−0.1151	−0.0335
p_value_nDCG	0.8054	0.9216	-	−0.2350	−0.2033	−0.0857
ER_P@10	0.0939	0.0674	0.0756	-	0.5651	0.3091
ER_AP	0.2232	0.2082	0.2473	0.5886	-	0.5298
ER_nDCG	0.1006	0.1167	0.1559	0.2220	0.4318	-

In Sum



Replicability

- Measures behave as expected and consistently
- RBO, RMSE, ER/DeltaRI provide insights at different levels

Reproducibility

- More challenging
- ER/DeltaRI also provide insights of overall effects
- Unpaired t-test might be too sensitive

Conclusion

Conclusion

Impact

- Replicability and reproducibility **measures** for system-oriented IR experiments
- Reproducibility-oriented **dataset**

Future directives

- Exploit other statistical measures
- Better understanding of how **user experience** is affected

References i

M. Baker.
1,500 scientists lift the lid on reproducibility.
Nature, 533:452–454, May 2016.

N. Ferro, N. Fuhr, M. Maistro, T. Sakai, and I. Soboroff.
Overview of CENTRE@CLEF 2019: Sequel in the Systematic Reproducibility Realm.
In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Tenth International Conference of the CLEF Association (CLEF 2019)*, pages 287–300. Lecture Notes in Computer Science (LNCS) 11696, Springer, Heidelberg, Germany, 2019.

References ii

N. Ferro, M. Maistro, T. Sakai, and I. Soboroff.
Overview of CENTRE@CLEF 2018: a First Tale in the Systematic Reproducibility Realm.
In P. Bellot, C. Trabelsi, J. Mothe, F. Murtagh, J.-Y. Nie, L. Soulier, E. SanJuan, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Nineth International Conference of the CLEF Association (CLEF 2018)*, pages 239–246. Lecture Notes in Computer Science (LNCS) 11018, Springer, Heidelberg, Germany, 2018.

M. R. Grossman and G. V. Cormack.
MRG_Uwaterloo and WaterlooCormack Participation in the TREC 2017 Common Core Track.
In *Proceedings of The Twenty-Sixth Text REtrieval Conference, TREC 2017, Gaithersburg, Maryland, USA, November 15-17, 2017*, 2017.

M. G. Kendall.
Rank correlation methods.
Griffin, Oxford, England, 1948.

References iii

T. Sakai, N. Ferro, I. Soboroff, Z. Zeng, P. Xiao, and M. Maistro.
Overview of the NTCIR-14 CENTRE Task.
In E. Ishita, N. Kando, M. P. Kato, and Y. Liu, editors, *Proc. 14th NTCIR Conference on Evaluation of Information Access Technologies*, pages 494–509. National Institute of Informatics, Tokyo, Japan, 2019.

W. Webber, A. Moffat, and J. Zobel.
A Similarity Measure for Indefinite Rankings.
ACM Transactions on Information Systems (TOIS), 4(28):20:1–20:38, November 2010.