Appendix to 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

This appendix reports additional tables and figures, showing the full set of experiments about replicability and reproducibility. The reported results are coherent with those presented in the main paper. In the following, we provide a short overview of the included tables and figures:

- Table 1 shows results of the *replicated* advanced a-run `WCrobust0405`. This table is referred to in the main paper at the beginning of *5.1 Validation of Measures*. It corresponds to Table 1 that shows results of the baseline b-run in the main paper. Analogously, the ARP[1] (including P@10, AP[2], nDCG[3]), correlation measures (Kendall's τ and RBO[4]), the RMSE[5] and p-values are reported.
- Table 2 shows results of the *reproduced* advanced a-run `WCrobust0405`. In this case, the ARP and p-values are reported. It complements Table 4 in the main paper.
- Table 3 shows results that are computed with an unpaired t-test between runs that are derived with the same system but on different collections.
- Figure 1 shows Kendall's τ and RMSE values computed at different cut-offs for the *replicated* advanced a-run `WCrobust0405`. It complements Figure 1 in the main paper that shows results of the baseline b-run.
- Figure 2 shows plots of the ER[6] against DeltaRI[7] for the *replicated* runs. These plots complement Figure 2 in the main paper. More specifically, the plots are based on run constellations with varied parametrization of the classifier. `rpl_tol` includes runs with different tolerance values for the stopping criterion. `rpl_C` includes runs with modified parameters of the regularization strength.
- Figure 3 shows plots of the ER against DeltaRI for the *reproduced* runs. These plots complement Figure 3 in the main paper. Analogously to the replicated runs, the plots show results of different classifier parametrizations.

[1] Average Retrieval Performance
[2] Average Precision
[3] Normalized Discounted Cumulated Gain
[4] Rank-Biased Overlap
[5] Root Mean Square Error
[6] Effect Ratio
[7] Delta Relative Improvement

Table 1: Replicability results for `WCrobust0405`: ARP, Kendall's τ, RMSE, and p values returned by the paired t-test. These results of the advanced a-run corresponds to Table 1 in the main paper that shows results of the baseline b-run.

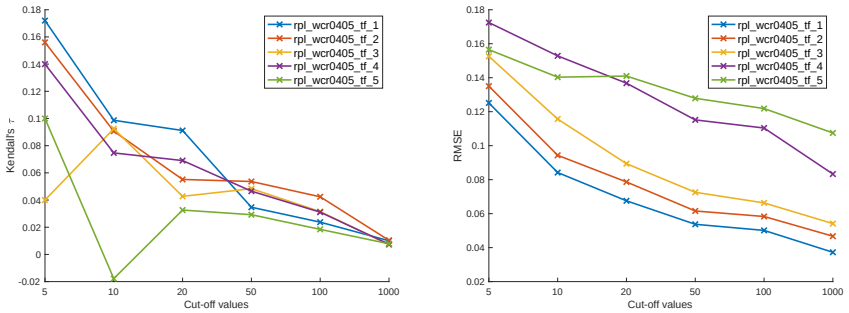
	ARP			Correlation		RMSE			p-value		
run	P@10	AP	nDCG	τ	RBO	P@10	AP	nDCG	P@10	AP	nDCG
WCrobust0405	0.7500	0.4278	0.6956	1	1	0	0	0	1	1	1
rpl_wcr0405_tf_1	0.7760	0.4233	0.6859	0.0100	0.6401	0.0927	0.0442	0.0373	0.046	0.470	0.063
rpl_wcr0405_tf_2	0.7660	0.4211	0.6841	0.0104	0.6133	0.0938	0.0510	0.0467	0.231	0.354	0.079
rpl_wcr0405_tf_3	0.7760	0.4186	0.6816	0.0071	0.5684	0.1122	0.0605	0.0541	0.101	0.287	0.066
rpl_wcr0405_tf_4	0.7340	0.3942	0.6631	0.0075	0.5304	0.1876	0.1002	0.0833	0.551	0.015	0.004
rpl_wcr0405_tf_5	0.7400	0.3711	0.6433	0.0078	0.4770	0.1913	0.1219	0.1075	0.715	$5E{-}04$	$2E{-}04$
rpl_wcr0405_df_1	0.7700	0.4136	0.6789	0.0153	0.6744	0.1020	0.0419	0.0373	0.167	0.014	$9E{-}04$
rpl_wcr0405_df_2	0.7620	0.3947	0.6663	0.0125	0.6573	0.1020	0.0530	0.0564	0.410	$9E{-}07$	$9E{-}05$
rpl_wcr0405_df_3	0.7100	0.3504	0.6286	0.0066	0.5561	0.1249	0.1008	0.1043	0.021	$4E{-}11$	$3E{-}07$
rpl_wcr0405_df_4	0.6220	0.2854	0.5570	0.0111	0.4793	0.2107	0.1729	0.1900	$2E{-}06$	$1E{-}13$	$1E{-}09$
rpl_wcr0405_df_5	0.5380	0.2320	0.4891	0.0085	0.3817	0.3105	0.2296	0.2668	$3E{-}08$	$1E{-}15$	$2E{-}11$
rpl_wcr0405_tol_1	0.7820	0.4161	0.6780	0.0096	0.6736	0.0980	0.0550	0.0451	0.019	0.132	0.004
rpl_wcr0405_tol_2	0.7060	0.3725	0.6031	0.0126	0.5890	0.2315	0.1455	0.2318	0.181	0.005	0.003
rpl_wcr0405_tol_3	0.5640	0.3031	0.4938	0.0068	0.4634	0.3947	0.2196	0.3445	$4E{-}04$	$1E{-}05$	$6E{-}06$
rpl_wcr0405_tol_4	0.4360	0.2175	0.3674	0.0039	0.3333	0.4930	0.3053	0.4610	$5E{-}07$	$2E{-}08$	$4E{-}09$
rpl_wcr0405_tol_5	0.2000	0.0682	0.1463	0.0013	0.1287	0.6479	0.4073	0.6001	$3E{-}15$	$1E{-}17$	$5E{-}21$
rpl_wcr0405_C_1	0.7680	0.4028	0.6713	0.0087	0.6648	0.0860	0.0540	0.0467	0.140	$6E{-}04$	$8E{-}05$
rpl_wcr0405_C_2	0.7800	0.4135	0.6786	0.0133	0.6934	0.0949	0.0434	0.0384	0.023	0.017	0.001
rpl_wcr0405_C_3	0.7740	0.4167	0.6802	0.0062	0.6605	0.0917	0.0514	0.0431	0.063	0.128	0.009
rpl_wcr0405_C_4	0.7200	0.3828	0.6518	0.0036	0.5571	0.1581	0.0903	0.0834	0.182	$1E{-}04$	$7E{-}05$
rpl_wcr0405_C_5	0.7060	0.3722	0.6424	0.0096	0.5279	0.1918	0.1047	0.0987	0.105	$5E{-}05$	$4E{-}05$

Table 2: Reproducibility: ARP and p-value (unpaired t-test), for `WCrobust0405`. Opposed to the replicability case, these runs rank documents of a different corpus than in the original setup. Thus, only ARP and p-values are reported. It complements table 4 in the main paper with results of the advanced a-run.

	ARP			p-value		
run	P@10	AP	nDCG	P@10	AP	nDCG
rpd_tf_1	0.4920	0.2341	0.5065	$3E{-}04$	$7E{-}06$	$9E{-}06$
rpd_tf_2	0.4760	0.2377	0.5090	$1E{-}04$	$9E{-}06$	$1E{-}05$
rpd_tf_3	0.4840	0.2354	0.5073	$2E{-}04$	$7E{-}06$	$1E{-}05$
rpd_tf_4	0.4520	0.2286	0.4943	$6E{-}05$	$5E{-}06$	$6E{-}06$
rpd_tf_5	0.4520	0.1993	0.4645	$3E{-}05$	$1E{-}07$	$1E{-}07$
rpd_df_1	0.4720	0.2294	0.5103	$1E{-}04$	$3E{-}06$	$1E{-}05$
rpd_df_2	0.4640	0.2252	0.5113	$7E{-}05$	$2E{-}06$	$8E{-}06$
rpd_df_3	0.4080	0.2066	0.4926	$3E{-}06$	$1E{-}07$	$1E{-}06$
rpd_df_4	0.3760	0.1750	0.4489	$3E{-}07$	$3E{-}09$	$2E{-}08$
rpd_df_5	0.3360	0.1416	0.3920	$2E{-}08$	$4E{-}11$	$4E{-}10$
rpd_tol_1	0.4800	0.2245	0.4984	$1E{-}04$	$2E{-}06$	$4E{-}06$
rpd_tol_2	0.4800	0.2064	0.4636	$2E{-}04$	$4E{-}07$	$7E{-}07$
rpd_tol_3	0.4120	0.1811	0.4075	$1E{-}05$	$2E{-}08$	$3E{-}08$
rpd_tol_4	0.3200	0.1389	0.2863	$1E{-}07$	$1E{-}09$	$3E{-}11$
rpd_tol_5	0.0480	0.0071	0.0376	$6E{-}21$	$3E{-}21$	$2E{-}34$
rpd_C_1	0.4840	0.2299	0.4999	$2E{-}04$	$3E{-}06$	$54E{-}06$
rpd_C_2	0.4920	0.2330	0.5052	$3E{-}04$	$5E{-}06$	$8E{-}06$
rpd_C_3	0.4840	0.2259	0.5013	$2E{-}04$	$3E{-}06$	$5E{-}06$
rpd_C_4	0.4280	0.2024	0.4704	$2E{-}05$	$3E{-}07$	$4E{-}07$
rpd_C_5	0.4120	0.1958	0.4597	$8E{-}06$	$1E{-}07$	$1E{-}07$

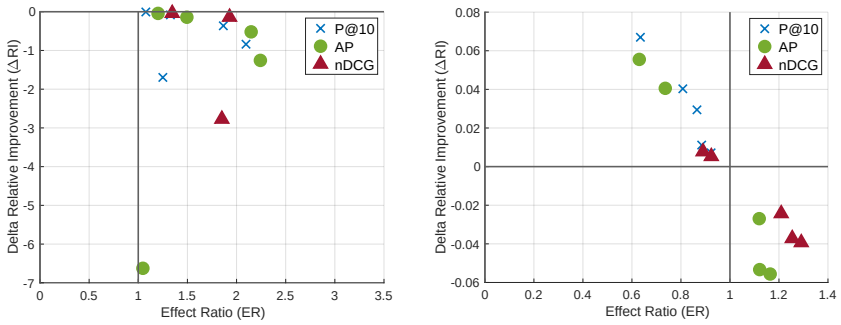
Table 3: Reproducibility: p-value (unpaired t-test), for `WCrobust04` and `WCrobust0405`. We compute an unpaired t-test between runs that are derived with the same system but on different collections, e.g. `rpl_wcrobust04_tf_1` with `rpd_wcrobust04_tf_1`. Most of the p-values are low, indicating that the two collections are quite different.

	WCrobust04			WCrobust0405		
run	P@10	AP	nDCG	P@10	AP	nDCG
rpd_tf_1	$8E{-}05$	$1E{-}05$	$7E{-}05$	$8E{-}05$	$2E{-}05$	$4E{-}05$
rpd_tf_2	$1E{-}04$	$2E{-}05$	$5E{-}05$	$8E{-}05$	$3E{-}05$	$7E{-}05$
rpd_tf_3	$1E{-}05$	$1E{-}05$	$3E{-}05$	$6E{-}05$	$4E{-}05$	$9E{-}05$
rpd_tf_4	$7E{-}06$	$8E{-}06$	$1E{-}05$	$2E{-}04$	$3E{-}04$	$3E{-}04$
rpd_tf_5	$7E{-}05$	$3E{-}05$	$2E{-}05$	$1E{-}04$	$1E{-}04$	$1E{-}04$
rpd_df_1	0.001	$3E{-}04$	0.002	$5E{-}05$	$2E{-}05$	$1E{-}04$
rpd_df_2	0.002	0.001	0.010	$3E{-}05$	$6E{-}05$	$3E{-}04$
rpd_df_3	0.007	0.003	0.013	$4E{-}05$	$4E{-}04$	0.002
rpd_df_4	0.017	0.013	0.073	0.001	0.006	0.031
rpd_df_5	0.055	0.017	0.094	0.010	0.021	0.084
rpd_tol_1	0.002	$5E{-}04$	0.006	$2E{-}05$	$2E{-}05$	$5E{-}05$
rpd_tol_2	0.065	0.028	0.291	0.007	$7E{-}04$	0.026
rpd_tol_3	0.064	0.097	0.246	0.116	0.025	0.263
rpd_tol_4	0.046	0.051	0.048	0.246	0.152	0.329
rpd_tol_5	0.098	0.094	0.031	0.025	0.053	0.030
rpd_C_1	$1E{-}07$	$7E{-}07$	$1E{-}07$	$8E{-}05$	$6E{-}05$	$9E{-}05$
rpd_C_2	$1E{-}07$	$3E{-}07$	$1E{-}07$	$5E{-}05$	$3E{-}05$	$7E{-}05$
rpd_C_3	$1E{-}07$	$6E{-}08$	$2E{-}08$	$4E{-}05$	$2E{-}05$	$6E{-}05$
rpd_C_4	$8E{-}04$	$4E{-}04$	0.003	$2E{-}04$	$8E{-}05$	$1E{-}04$
rpd_C_5	0.001	$5E{-}04$	0.003	$4E{-}04$	$1E{-}04$	$2E{-}04$



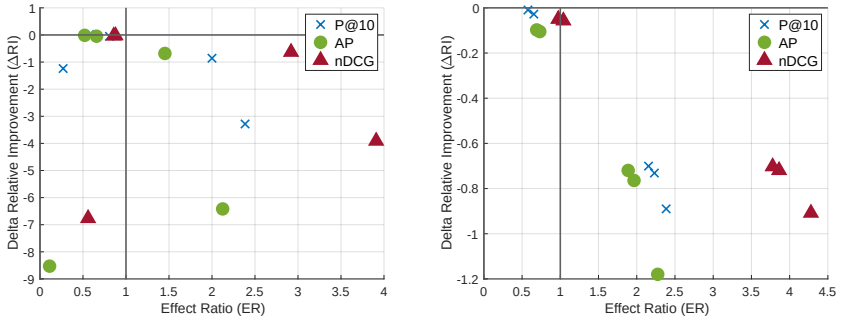
(a) Kendall's τ `WCrobust0405`. (b) RMSE `WCrobust0405`.

Figure 1: Kendall's τ and RMSE values computed at different cut-offs for replicated `WCrobust0405` runs. The plots complement Figure 1 in the main paper with results of the advanced a-run.



(a) `rpl_tol` runs. (b) `rpl_C` runs.

Figure 2: Replicability: ER on the x-axis against DeltaRI on the y-axis. These plots complement Figure 2 in the main paper. More specifically, the plots are based on run constellations with varied parametrization of the classifier. `rpl_tol` varies tolerance values of the stopping criterion. `rpl_C` varies the ℓ^2 regularization strength.



(a) `rpd_tol` runs. (b) `rpd_C` runs.

Figure 3: Reproducibility: ER on the x-axis against DeltaRI on the y-axis. These plots complement Figure 3 in the main paper. More specifically, the plots are based on run constellations with varied parametrization of the classifier. Analogously to the replicated runs, `rpd_tol` and `rpd_C` vary the tolerance values for the stopping criterion and the ℓ^2 regularization strength, respectively.