Beel, Joeran. A Call for Evidence-based Best-Practices for Recommender Systems Evaluations. 2024. In Report from Dagstuhl Seminar 24211: Evaluation Perspectives of Recommender Systems: Driving Research and Education. Editors: Bauer, Christine / Said, Alan / Zangerle, Eva.

A Call for Evidence-based Best-Practices for Recommender Systems Evaluations

Joeran Beel (University of Siegen / Recommender-Systems.com – Siegen, Germany, joeran.beel@uni-siegen.de)

I recall vividly when more than a decade ago – I was a PhD student – *Konstan & Adomavicius* warned that *"the recommender systems research community [...] is facing a crisis where a significant number of research papers lack the rigor and evaluation to be properly judged and, therefore, have little to contribute to collective knowledge* [14]". Similar concerns were already voiced two years earlier by Ekstrand et al. [12]. Over the following years, many more researchers criticized the evaluation practices in the community [13, 21, 19, 10], myself included [5, 8, 4, 20, 23, 15, 6, 7]. The situation may have somewhat improved in the past years due to more awareness in the community [13], the reproducibility track at the ACM RecSys conference, innovative submission formats like "result-blind reviews" [9] via registered reports at ACM TORS, and several new software libraries, including Elliot [1], RecPack [16], Recbole [25], and LensKit-Auto [22]. Yet the decade-old criticism by *Konstan & Adomavicius* remains as true today as it was a decade ago.

Konstan & Adomavicius proposed that, among others, best-practice guidelines on recommender systems research and evaluations might offer a solution to the crisis [14]. In their paper, they also presented results from a small survey that indicated that such guidelines would be welcomed by many members of the community. However, to my knowledge, no comprehensive guidelines or checklists have been specifically created for the recommender systems community, or at least they have not been widely adopted. Recently, I attempted to develop guidelines for releasing recommender systems research code [3], based on the NeurIPS and 'Papers with Code' guidelines [24], but progress has been limited.

I echo the demand[1] by *Konstan & Adomavicius* [14] for the recommender systems community to establish best-practice guidelines and/or checklists for researchers and reviewers. Such guidelines would facilitate the conduct of 'good' research, and they would assist reviewers in conducting through reviews. By 'good research' I primarily mean reproducible research with a sound methodology. But 'good' research also refers to research that others easily can build upon, e.g. because data and code are available; research that is ethical; and research that is sustainable, e.g. because no resources were wasted.

My vision is best-practice guidelines that are not merely a collection of opinions but are instead grounded in empirical evidence. This approach would be analogous to the medical field, where guidelines for practitioners are justified based on empirical research findings. Additionally, these medical guidelines indicate the degree of consensus among experts, allowing medical practitioners to understand how widely accepted each best practice is. In areas with less expert consensus, deviations from the best practice by practitioners would be more acceptable. This model ensures that guidelines are both scientifically robust and flexible.

In my view, best-practice guidelines for recommender systems research and evaluation should include the following components in addition to the best practices themselves:

[1] Please note that I used ChatGPT to improve my writing. I wrote all the sentences first myself and then asked ChatGPT for each paragraph to improve the writing but keep the structure.

1. Justification: A justification for the best practice, ideally based on empirical evidence.
2. Confidence: An estimate of how sound the evidence is.
3. Severity: An estimate of the importance of the best practice and the potential consequences of not following it.
4. Consensus: The degree of agreement within the community or among experts that the proposed best practice is indeed a best practice.

Table 1 illustrates what a best practice may look like, using the example of random seeds. A random seed is an initial value for a pseudo-random number generator, ensuring that the sequence of random numbers it produces is reproducible. This reproducibility is crucial for consistent experiment results, fair comparisons between different algorithms, and reliable debugging. For instance, when splitting a dataset into training and testing sets, using a fixed random seed ensures the same split is produced each time. This consistency allows researchers to compare the performance of different algorithms on identical data splits, ensuring that any performance differences are due to the algorithms themselves and not variations in the data splits. Generating random random-seeds is not a trivial task, and dedicated tools exist for it [11].

Creating a preliminary set of guidelines for recommender systems evaluation should be straightforward. Existing communities, particularly in machine learning, already have robust best-practice guidelines and checklists. Notably, NeurIPS [17, 18] and the AutoML conference [2] offer guidelines that could be adapted for recommender system experiments with relatively minor modifications. Initially, these guidelines do not require empirical evidence or consensus surveys. They can be simple and aligned with those used in the machine-learning community. Over time, these guidelines can be tailored more to fit recommender systems research, expanded and substantiated with empirical evidence and broader consensus.

The creation and justification of best practices can likely be undertaken by any motivated researcher with experience in recommender systems research. However, the final selection of these best practices, particularly concerning points 3 (severity) and 4 (consensus), should be conducted by reputable members of the RecSys community. This could be achieved through a Dagstuhl seminar with selected experts or by the steering committee of the ACM Recommender Systems Conference.

In conclusion, establishing well-defined best-practice guidelines, endorsed by the community and enforced by key publication venues such as the ACM Recommender Systems conference and the ACM Transactions on Recommender Systems (TORS) journal, would be a significant move towards resolving the long-standing crisis in the recommender system research community. For over a decade, the community has struggled with inconsistencies and lack of rigor in research practices. By adopting and enforcing these guidelines, we can ensure higher research standards, facilitate reproducibility, and contribute more robustly to collective knowledge.

Random Seeds Best-Practice	1) Experiments must be repeated (n>=5) with different random seeds each time. This is true for each aspect of an experiment that requires randomness. This includes splitting data and initializing weights in neural networks. 2) The exact random seeds used for experiments must be reported in the paper or the code.
Justification	In the context of data splitting, Wegmeth et al. [23] showed that when random seeds differed – i.e. data splits contained different data due to randomness – the performance of the same algorithm, with the same hyper-parameters varied by up to 12% [23]. In contrast, repeating and averaging experiments with different random seeds, led to a maximum difference of only around 4%. This means, if only a single run had been conducted, the results could be up to 6% above or under the 'true' result, possibly more. By repeating the experiments, the difference would have been only ±2% in the worst case. The variance depended on the applied metrics, cut-offs, datasets, and splitting methods (lower variance for cross-fold validation, higher variance for hold-out validation). Therefore, repeating experiments with different random seeds ensures that the reported result is closer to the 'true' result. Reporting the exact random seeds is also a prerequisite (besides many other factors) for an exact replication of experiments. A researcher who wants to replicate an experiment and who uses the identical random seeds as the original researcher, will have the same data in the train and validation splits as the original researcher. Knowing the exact random seeds also makes it easier to detect fraudulent behavior such as cherry picking.
Severity	Medium: If not conducted properly, reported results may be off the 'true' results by multiple per cent.
Confidence	Low (the empirical evidence is based only on one workshop publication [23]).
Consensus	82% of the ACM RecSys Steering Committee agree with this best practice. *PLEASE NOTE: This is an example for illustration purposes. The percentage is made up.*

Table 1 Best Practices for Random Seeds (Example)

References

1 Vito Walter Anelli, Alejandro Bellogin, Antonio Ferrara, Daniele Malitesta, Felice Antonio Merra, Claudio Pomo, Francesco Maria Donini, and Tommaso Di Noia. Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2405–2414, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450380379. doi: 10.1145/3404835.3463245. URL https://doi.org/10.1145/3404835.3463245.

2 AutoML. Automl author instructions. *https://github.com/automl-conf/LatexTemplate/blob/main/instructions.pdf*, 2024.

3 Joeran Beel. Releasing recsys research code. *https://github.com/ISG-Siegen/Releasing-RecSys-Research-Code*, 2023.

4 Joeran Beel and Victor Brunel. Data pruning in recommender systems research: Best-practice or malpractice? In *13th ACM Conference on Recommender Systems (RecSys)*, volume 2431, pages 26–30. CEUR-WS, 2019.

5 Joeran Beel and Stefan Langer. A comparison of offline evaluations, online evaluations, and user studies in the context of research-paper recommender systems. In Sarantos Kapidakis, Cezary Mazurek, and Marcin Werla, editors, *Proceedings of the 19th International Conference on Theory and Practice of Digital Libraries (TPDL)*, volume 9316 of *Lecture Notes in Computer Science*, pages 153–168, 2015. doi: 10.1007/978-3-319-24592-8_12.

6 Joeran Beel, Stefan Langer, Marcel Genzmehr, Bela Gipp, and Andreas Nürnberger. A comparative analysis of offline and online evaluations and discussion of research paper recommender system evaluation. In *Proceedings of the Workshop on Reproducibility and Replication in Recommender Systems Evaluation (RepSys) at the ACM Recommender System Conference (RecSys)*, ACM International Conference Proceedings Series (ICPS), pages 7–14, 2013. doi: 10.1145/2532508.2532511.

7 Joeran Beel, Stefan Langer, Andreas Nuenberger, and Marcel Genzmehr. The impact of demographics (age and gender) and other user characteristics on evaluating recommender systems. In Trond Aalberg, Milena Dobreva, Christos Papatheodorou, Giannis Tsakonas, and Charles Farrugia, editors, *Proceedings of the 17th International Conference on Theory and Practice of Digital Libraries (TPDL 2013)*, pages 400–404, Valletta, Malta, September 2013. Springer.

8 Joeran Beel, Corinna Breitinger, Stefan Langer, Andreas Lommatzsch, and Bela Gipp. Towards reproducibility in recommender-systems research. *User Modeling and User-Adapted Interaction (UMUAI)*, 26(1):69–101, 2016. doi: 10.1007/s11257-016-9174-x.

9 Joeran Beel, Timo Breuer, Anita Crescenzi, Norbert Fuhr, and Meije Li. Results-blind reviewing. *Dagstuhl Reports*, 13(1):68–154, 2023. doi: 10.4230/DagRep.13.1.68. URL https://drops.dagstuhl.de/opus/volltexte/2023/19119/pdf/dagrep_v013_i001_p068_23031.pdf.

10 Alejandro Bellogín and Alan Said. Improving accountability in recommender systems research through reproducibility. *User Model. User Adapt. Interact.*, 31(5): 941–977, 2021. doi: 10.1007/S11257-021-09302-X. URL https://doi.org/10.1007/s11257-021-09302-x.

11 Michael D. Ekstrand. seedbank: easy management of seeds across RNGs, May 2024. URL https://doi.org/10.5281/zenodo.11276061.

12 Michael D. Ekstrand, Michael Ludwig, Joseph A. Konstan, and John T. Riedl. Rethinking the recommender research ecosystem: reproducibility, openness, and LensKit. In *Proceedings of the Fifth ACM Conference on Recommender Systems*, RecSys '11, page 133–140,

REFERENCES

New York, NY, USA, 2011. Association for Computing Machinery. ISBN 9781450306836. doi: 10.1145/2043932.2043958. URL https://doi.org/10.1145/2043932.2043958.

13 Maurizio Ferrari Dacrema, Paolo Cremonesi, and Dietmar Jannach. Are we really making much progress? a worrying analysis of recent neural recommendation approaches. In *Proceedings of the 13th ACM Conference on Recommender Systems*, RecSys '19, page 101–109, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450362436. doi: 10.1145/3298689.3347058. URL https://doi.org/10.1145/3298689.3347058.

14 Joseph A. Konstan and Gediminas Adomavicius. Toward identification and adoption of best practices in algorithmic recommender systems research. In *Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation*, RepSys '13, page 23–28, New York, NY, USA, 2013. Association for Computing Machinery. ISBN 9781450324656. doi: 10.1145/2532508.2532513. URL https://doi.org/10.1145/2532508.2532513.

15 Stefan Langer and Joeran Beel. The comparability of recommender system evaluations and characteristics of docear's users. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (REDD) at the 2014 ACM Conference Series on Recommender Systems (RecSys)*, pages 1–6. CEUR-WS, 2014.

16 Lien Michiels, Robin Verachtert, and Bart Goethals. Recpack: An(other) experimentation toolkit for top-n recommendation using implicit feedback data. In *Proceedings of the 16th ACM Conference on Recommender Systems*, RecSys '22, page 648–651, New York, NY, USA, 2022. Association for Computing Machinery. ISBN 9781450392785. doi: 10.1145/3523227.3551472. URL https://doi.org/10.1145/3523227.3551472.

17 NeurIPS. Neurips 2022 paper checklist guidelines. *https://neurips.cc/Conferences/2022/PaperInformation/PaperChecklist*, 2022.

18 NeurIPS. Neurips paper checklist guidelines. *https://neurips.cc/public/guides/PaperChecklist*, 2024.

19 Alan Said and Alejandro Bellogín. Replicable evaluation of recommender systems. In Hannes Werthner, Markus Zanker, Jennifer Golbeck, and Giovanni Semeraro, editors, *Proceedings of the 9th ACM Conference on Recommender Systems, RecSys 2015, Vienna, Austria, September 16-20, 2015*, pages 363–364. ACM, 2015. URL https://dl.acm.org/citation.cfm?id=2792841.

20 Teresa Scheidt and Joeran Beel. Time-dependent evaluation of recommender systems. In *Perspectives on the Evaluation of Recommender Systems Workshop, ACM RecSys Conference*, 2021. URL https://ceur-ws.org/Vol-2955/paper10.pdf.

21 Faisal Shehzad and Dietmar Jannach. Everyone's a winner! on hyperparameter tuning of recommendation models. In *17th ACM Conference on Recommender Systems*, 2023.

22 Tobias Vente, Michael Ekstrand, and Joeran Beel. Introducing LensKit-Auto, an experimental automated recommender system (autorecsys) toolkit. In *Proceedings of the 17th ACM Conference on Recommender Systems*, pages 1212–1216, 2023. URL https://dl.acm.org/doi/10.1145/3604915.3610656.

23 Lukas Wegmeth, Tobias Vente, Lennart Purucker, and Joeran Beel. The effect of random seeds for data splitting on recommendation accuracy. In *Proceedings of the 3rd Perspectives on the Evaluation of Recommender Systems Workshop*, 2023. URL https://ceur-ws.org/Vol-3476/paper4.pdf.

24 Papers with code. Code template. *https://github.com/paperswithcode/releasing-research-code/blob/master/templates/README.md*, 2000.

25 Wayne Xin Zhao, Shanlei Mu, Yupeng Hou, Zihan Lin, Yushuo Chen, Xingyu Pan, Kaiyuan Li, Yujie Lu, Hui Wang, Changxin Tian, Yingqian Min, Zhichao Feng, Xinyan

REFERENCES

Fan, Xu Chen, Pengfei Wang, Wendi Ji, Yaliang Li, Xiaoling Wang, and Ji-Rong Wen. Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4653–4664, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450384469. doi: 10.1145/3459637.3482016. URL https://doi.org/10.1145/3459637.3482016.