Exploring Effect-Size-Based Meta-Analysis for Multi-Dataset Evaluation

Mete Sertkan[1], Sophia Althammer[2], Sebastian Hofstätter[3], Peter Knees[2] and Julia Neidhardt[1]

[1]Christian Doppler Laboratory for Recommender Systems, TU Wien, Vienna, Austria
[2]TU Wien, Vienna, Austria
[3]Cohere, Vienna, Austria

Abstract

In this paper, we address the essential yet complex task of evaluating Recommender Systems (RecSys) across multiple datasets. This is critical for gauging their overall performance and applicability in various contexts. Owing to the unique characteristics of each dataset and the variability in algorithm performance, we propose the adoption of effect-size-based meta-analysis, a proven tool in comparative research. This approach enables us to compare a "treatment model" and a "control model" across multiple datasets, offering a comprehensive evaluation of their performance. Through two case studies, we highlight the flexibility and effectiveness of this method in multi-dataset evaluations, irrespective of the metric utilized. The power of forest plots in providing an intuitive and concise summarization of our analysis is also demonstrated, which significantly aids in the communication of research findings. Our work provides valuable insights into leveraging these methodologies to draw more reliable and validated conclusions on the generalizability and robustness of RecSys models.

Keywords

recommender systems, evaluation, effect-size, meta-analysis, forest plots,

1. Introduction

Within the fast-evolving domain of Recommender Systems (RecSys), broad adoption across various industries has prompted researchers to strive for improvements in general-purpose methods. Often, however, the effectiveness of improvements introduced by these novel methods is confined to highly specific experimental settings, encompassing particular datasets, evaluation measures, and baselines. As such, these enhancements do not necessarily translate into broad applicability across different contexts or problem domains [1]. Therefore, it's absolutely critical to evaluate recommender systems over multiple datasets to gain a more comprehensive understanding of their robustness and generalizability.

Evaluating recommender systems using multiple datasets is a complex process due to each dataset's unique characteristics and the algorithms' variability. The performance of an algorithm

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2023), September 19th, 2023, co-located with the 17th ACM Conference on Recommender Systems, Singapore, Singapore.
mete.sertkan@tuwien.ac.at (M. Sertkan); sophia.althammer@tuwien.ac.at (S. Althammer); sebastian.hofstaetter@tuwien.ac.at (S. Hofstätter); peter.knees@tuwien.ac.at (P. Knees); julia.neidhardt@tuwien.ac.at (J. Neidhardt)
0000-0003-0984-5221 (M. Sertkan); 0000-0001-7184-1841 (J. Neidhardt)

can vary greatly, excelling in one dataset while falling short in another, particularly when there are significant differences in the properties of the datasets, such as the ratio of users to items or the density of ratings [2, 3]. This variation in performance complicates drawing reliable conclusions, introducing potential subjective biases and the problematic comparison or aggregation of incompatible metrics or scenarios. Meanwhile, disciplines like social and medical sciences have already established robust tools and norms for such meta-analysis [4, 5]. Researchers in Information Retrieval (IR) and Natural Language Processing (NLP) have started to adopt these methodologies for multi-task evaluation [6, 5], underscoring the necessity and significance of such an approach for recommender systems.

In this study, we examine the value of effect-size-based meta-analysis as a tool for comparative research [4]. This method allows us to assess the implications of adopting a "treatment model", which could be a novel update or unique architectural design, compared to a "control model", possibly a baseline strategy or the current state-of-the-art. By using this technique across multiple datasets, we can better understand the broader performance and generalizability of these models. Furthermore, effect-size-based meta-analysis not only examines the impact on individual datasets but also consolidates the effects across various datasets into a unified statistical evaluation. This provides a reliable measure of the model's capacity to generalize across different datasets and reveals the significance or contribution of each dataset to the overall effect. It is important to note that this analytical method requires the availability of pairwise metrics - for treatment and control - for each sample in the datasets under consideration. However, it is versatile and not limited to any specific metric and can accommodate experiments that provide a variety of metrics, including accuracy or beyond-accuracy metrics. The results of the effect-size-based meta-analysis can be concisely visualized using forest plots [7], enhancing the interpretation and communication of the findings (refer to Figure 1).

In summary, this paper makes the following key contributions:

- We propose the use of effect-size-based meta-analysis as a robust approach for multi-dataset evaluations.
- We demonstrate the practical utility of this approach through two case studies, one involving an incremental treatment model and the other involving a more sophisticated update.
- We provide code and data of our experiments publicy available under: https://github.com/MeteSertkan/meta-analysis-based-recsys-eval

2. Related Work

Evaluating recommender systems poses inherent challenges, including varying algorithm performance across different datasets, divergent evaluation goals, and the complexity of choosing appropriate metrics for comparison [2, 3]. Among the three main evaluation types - offline experiments, user studies, and online evaluation - offline experiments are commonly preferred due to cost efficiency [8, 9, 10, 11]. However, the risk lies in using selectively curated datasets to demonstrate improvements, which may overemphasize the importance of quantitative measures in offline experimentation and distort the actual impact of our research [1, 3].

Yet, offline experiments still serve as a crucial initial step towards comprehensive evaluation. Several toolkits like RecPack [12], Elliot [13], Cornac [14], and RecBole [15] have been introduced for reproducible experimentation and evaluation. These tools provide the capabilities for easy, reproducible experiments, offer built-in baselines, models, data, and facilitate the evaluation and comparison of models using various metrics. Evaluating recommender systems over multiple datasets is essential for a comprehensive understanding of their robustness and generalizability. However, the unique characteristics of datasets, metrics, and performance variations across datasets complicate drawing reliable conclusions and can introduce subjective biases. For example, it's not valid to average *NDCG* scores across multiple datasets as *NDCG* scores are task-dependent and can only be compared within one task.

Despite their invaluable contributions, these tools do not fully address the need for statistically robust comparison and aggregation methods across diverse datasets and scenarios. This work proposes the use of effect-size-based meta-analysis [4] for this purpose, and we demonstrate its utility through two use cases.

3. Methods

We utilize effect-size-based meta-analysis to contrast the efficacy of a treatment model with a control model across multiple datasets. The treatment model could be an updated version of the control model, a new model, or one trained with additional data, while the control model acts as the standard for comparison. We consider the raw mean difference D and the standardized mean difference d, as defined by Borenstein et al. [4] and implemented in Ranger [5], to compute the effect-sizes and, in turn, the summary effect.

Raw Mean Difference *D.* In RecSys experiments, performance metrics are typically calculated for each user, item, or session. Averaging the metrics enable researchers to compare the effectiveness of different models. Thus, the mean difference, a direct and intuitive measurement of effect-size, aligns with the scale of the underlying metric. We compute the raw mean difference D by averaging the pairwise differences between treatment X_T and control metric X_C and use the standard deviation (S_{diff}) of the pairwise differences to compute its corresponding variance V_D as follows (n is the number of compared pairs):

$$D = \frac{X_T - X_C}{n}, \quad V_D = \frac{S_{\text{diff}}^2}{n}, \tag{1}$$

Standardized Mean Difference *d.* We might consider standardizing the mean difference (i.e., convert it into a "unitless" form) to make the effect-size comparable and combinable across studies, for example, in case of *RMSE* and different ratings scales. The standardized mean difference d is computed by dividing the raw mean difference D by the within-group standard deviation S_{within} calculated across the treatment and control metrics.

$$d = \frac{D}{S_{\text{within}}} \tag{2}$$

S_{within} is determined by the standard deviation of the pairwise differences S_{diff} and the correlation

of the corresponding pairs r as follows:

$$S_{\text{within}} = \frac{S_{\text{diff}}}{\sqrt{2(1-r)}} \tag{3}$$

The variance of standardized mean difference d is

$$V_d = (\frac{1}{n} + \frac{d^2}{2n})2(1-r), \tag{4}$$

where n is the number of compared pairs. In small samples, d tends to overestimate the absolute value of the true standardized mean difference δ, which can be corrected by factor J to obtain an unbiased estimate called Hedges' g [16, 4] and its corresponding variance V_g:

$$J = 1 - \frac{3}{4df - 1}, \quad g = J \times d, \quad V_g = J^2 \times V_d, \tag{5}$$

where df is degrees of freedom which is $n-1$ in the paired study setting with n number of pairs.

Combined Effect M^*. After calculating the individual effect-sizes (Y_i) and corresponding variances (V_{Y_i}) for k experiments (i.e., datasets), the final step is to synthesize them into one combined effect. We assume, as in [6, 5], that the effect-size variance varies across the used datasets, i.e., heterogeneity. Therefore, we employ the random-effects model as defined in [4] to consider the between-study variance T^2 for the summary effect computation. We use the DerSimonian and Laird method [17] to estimate T^2:

$$\begin{aligned} T^2 &= \frac{Q - df}{C}, \\ Q &= \sum_{i=1}^{k} W_i Y_i^2 - \frac{(\sum_{i=1}^{k} W_i Y_i^2)^2}{\sum_{i=1}^{k} W_i}, \\ df &= k - 1, \\ C &= \sum W_i - \frac{\sum W_i^2}{\sum W_i}. \end{aligned} \tag{6}$$

where the weight of the individual experiments $W_i = 1/V_{Y_i}$. We adjust the weights by T^2 and compute the weighted average of the individual effect-sizes, i.e., the summary effect M^*, and its corresponding variance V_{M^*} as follows:

$$W_i^* = \frac{1}{V_{Y_i} + T^2}, \quad M^* = \frac{\sum_{i=1}^{k} W_i^* Y_i}{\sum_{i=1}^{k} W_i^*}, \quad V_{M^*} = \frac{1}{\sum_{i=1}^{k} W_i^*}. \tag{7}$$

Confidence Interval (CI) We determine the corresponding confidence interval (represented by the lower limit, LL_Y, and the upper limit, UL_Y) for a given effect-size Y, which can be the result of an individual experiment (Y_i) or the summary effect (M^*), as follows:

$$SE_Y = \sqrt{V_Y}, \quad LL_Y = Y - Z^\alpha \times SE_Y, \quad UL_Y = Y + Z^\alpha \times SE_Y, \tag{8}$$

where SE_Y is the standard error, V_Y the variance of the effect-size, and Z^α the Z-value corresponding to the desired significance level α. Given α we compute $Z^\alpha = ppf(1 - \frac{\alpha}{2})$, where $ppf()$ is the percent point function (we use scipy.stats.norm.ppf[1]). For example, $\alpha = 0.05$ yields the

[1] https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.norm.html

Table 1
Considered explicit-feedback datasets (as provided by Cornac [14]).

Dataset	#Users	#Items	#Interactions
amazon-clothing	5,377	3,393	13,689
amazon-digital-music	5,541	3,568	64,706
amazon-office	3,703	6,523	53,282
amazon-toy	19,412	11,924	167,597
filmtrust	1,508	2,071	35,497
movielens-10M	69,878	10,677	10,000,054
netflix-small	10,000	5,000	607,803

95% CI of $Y \pm 1.96 \times SE_Y$.

Forest Plots. Meta-analysis based on effect-sizes yields individual and combined effects from experiments, their confidence intervals, and weights that show each experiment's contribution to the combined effect. Forest plots [7] conveniently summarize these results, allowing for intuitive interpretation and easy communication of findings. Please refer to Figure 1 for an example. Effect-sizes and confidence intervals are represented as diamonds with whiskers $\vdash \blacklozenge \dashv$. The size of the diamonds corresponds to the weight of the experiments (W_i^*). The dotted line at zero denotes the absence of an effect. Suppose the confidence interval of an observed effect-size crosses this line. In that case, it indicates that the effect-size is not significant at the given confidence level, meaning that the effect is not detectable.

4. Experimental Setting

For the effect-size-based meta-analysis, we compute pairwise performance metrics for the models under comparison on a user basis. Therefore we utilize models and data from the Cornac framework [14]. We use the explicit-feedback datasets listed in Table 1. We split the data in each experiment with 80% for training and 20% for testing. To conduct the effect-size-based meta-analysis, we utilize Ranger [5]. We illustrate the utility of effect-size-based meta-analysis through two use cases:

1) We compare matrix factorization with (MF-bias) and without bias terms (MF) [18] to demonstrate an incremental update. Our meta-analysis is based on *RMSE* and *NDCG*@10, highlighting the differences in matrix completion and ranking tasks. For *RMSE*, we use the standardized mean difference as the effect-size index, accounting for varying scales in each experiment. For *NDCG*@10, we utilize the raw mean difference.

2) We compare matrix factorization (MF) [18] with Bayesian probabilistic ranking (BPR) [19] to illustrate a more sophisticated treatment focusing on ranking. We perform the meta-analysis across all datasets, movie-only datasets, and retail-only datasets, offering insights into overall and domain-specific effects. *NDCG*@10 is the base metric in each case, and we use the raw mean difference as our effect-size index.

We use the default (hyper)parameters as provided by Cornac [14]. Note that our study focuses on evaluation rather than the models themselves. For replicability, we provide all runs, i.e.,

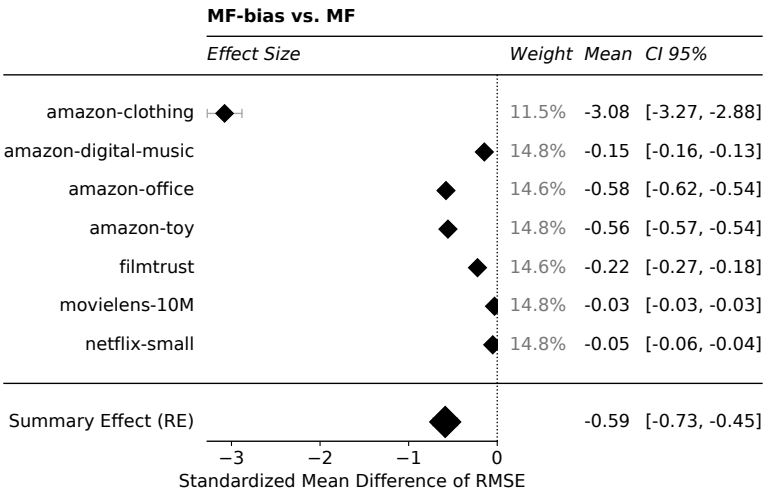

Figure 1: Comparing MF-bias (treatment) to MF (control) in terms of standardized mean difference of RMSE in matrix completion.

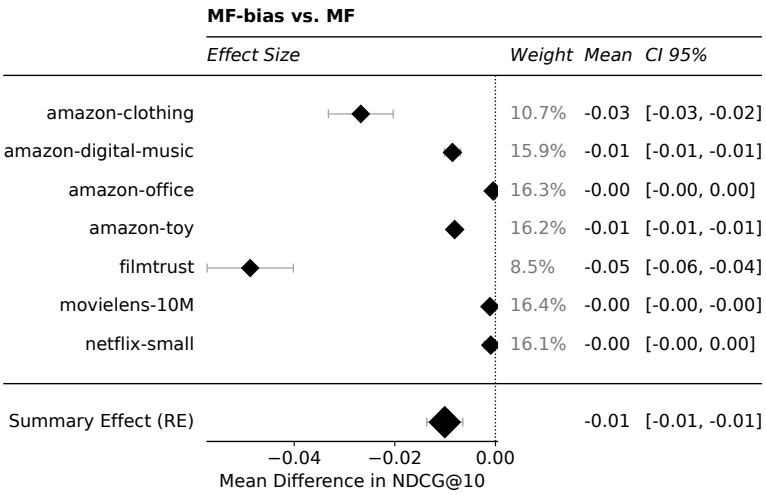

Figure 2: Comparing MF-bias (treatment) to MF (control) in terms of mean difference in *NDCG*@10 (ranking performance).

model-dataset-metric combinations.

5. Case Study MF-Bias vs. MF

In our first case study, we examine the impact of incorporating bias terms (user, item, and global) - designated as the treatment - against a basic matrix factorization model - the control. We measure their performance on user-item-rating matrix completion via *RMSE* and ranking through *NDCG*@10.

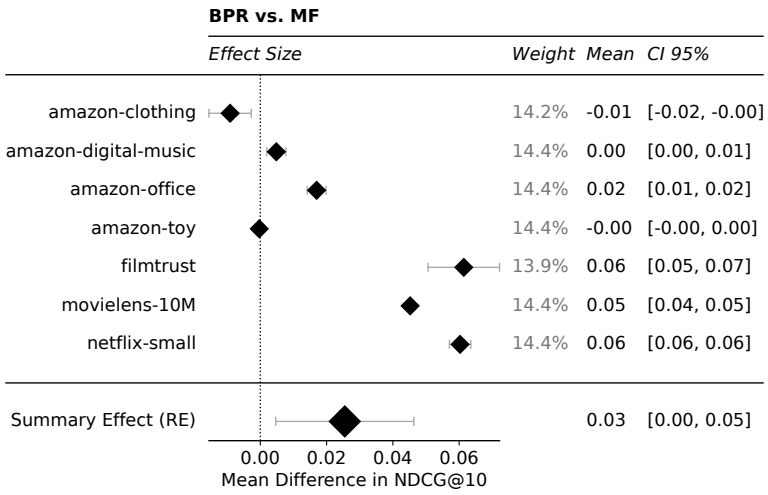

Figure 3: Comparing BPR (treatment) to MF (control) in terms of mean difference in NDCG@10 - All datasets.

Figure 1 summarizes the *RMSE* comparison. All confidence intervals (CI) lie on the left of the zero-effect (dotted) line, with none crossing it, suggesting that the use of bias terms consistently and significantly improves matrix completion performance (smaller error). Notably, nearly all experiments contribute similarly to the overall effect calculation, except for the *amazon-clothing* dataset, which shows a less confident effect estimation. This reflects the inverse relationship between the variance of an experiment's effect-size and its weight in calculating the summary effect, as detailed in Equations 6 and 7.

Regarding ranking performance (refer to Figure 2), we generally observe a decline when bias terms are introduced, suggesting potential overfitting. While introducing bias terms enhance matrix completion, it seems to negatively impact ranking performance. Notably, for individual experiments using *amazon-office* and *netflix-small* datasets, we find no significant ranking performance differences, as their corresponding confidence intervals cross the zero-effect line.

6. Case Study BPR vs. MF

In our second case study, we evaluate the effect of using Bayesian probabilistic ranking (BPR) - treatment - versus matrix factorization (MF) - control - on ranking performance ($NDCG@10$). This comparison is conducted across all datasets, as well as exclusively on movie and retail datasets. Looking at the summary effect in Figure 3, we generally expect a boost in ranking performance when choosing BPR over MF. Although the summary effect estimate is less certain than in the prior meta-analysis, it is still significant. The difference in ranking performance between individual experiments with retail and movie datasets is immediately noticeable. Focusing on retail data only (Figure 4), we expect a performance decline for *amazon-clothing*, no significant difference for *amazon-toy*, and performance gains for *amazon-digital-music* and *amazon-office*. On the whole, we do not expect significant differences when choosing BPR over MF for retail datasets. In contrast, for movie datasets, we expect consistent and significant

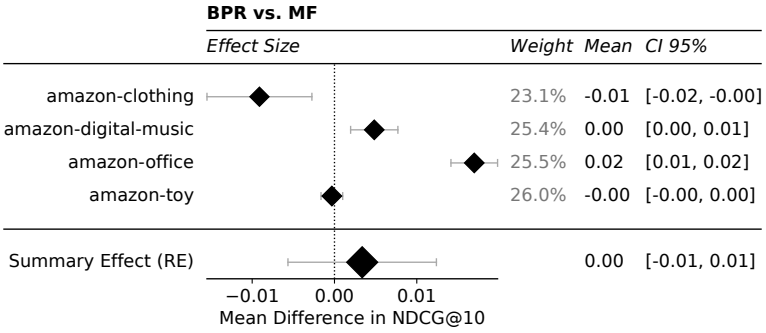

Figure 4: Comparing BPR (treatment) to MF (control) in terms of mean difference in NDCG@10 - Only retail datasets.

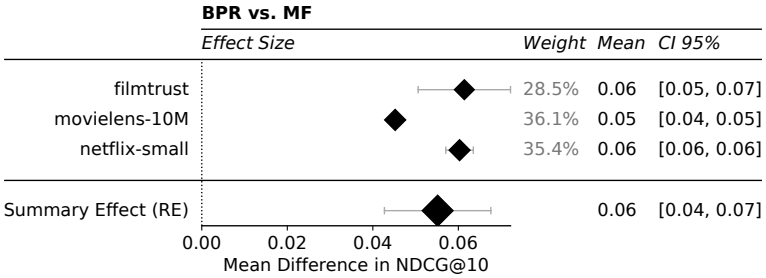

Figure 5: Comparing BPR (treatment) to MF (control) in terms of mean difference in NDCG@10 - Only movie datasets.

improvements in ranking performance when employing BPR (Figure 5).

7. Conclusions

Evaluating recommender systems across multiple datasets is crucial for understanding their generalizability and robustness. Yet, given the unique characteristics of different data and algorithms, making comparisons across datasets can be challenging. For instance, it's not straightforward to compare RMSE values or average NDCG values across multiple datasets to determine a model's overall capacity.

This is where effect-size-based meta-analysis comes in handy. It allows for a statistically robust comparison between two models across multiple datasets, providing a reliable synthesis of results. This mitigates subjective interpretations and fosters more valid conclusions on the overall treatment effect.

We have outlined the theoretical underpinning of this method and, through two case studies, demonstrated its utility in multi-dataset evaluations. This method isn't metric-dependent and is applicable even when scales vary across datasets. Unlike a p-value from a hypothesis test, which indicates the likelihood of correctly rejecting a null hypothesis (a retrospective view), confidence intervals predict future outcomes in similar experiments. They enable distinguishing between the magnitude of an effect and the probability of its recurrence [6]. Forest plots succinctly

summarize the analysis outcomes, enabling intuitive interpretation and facilitating research communication. For instance, they clearly highlighted the discrepancy between retail and movie domains in our second case study, which can guide dataset selection for evaluation - a significant challenge in itself [3].

Since recommender system evaluations often involve more than two models, our future work will adapt this approach for multi-dataset, multi-model settings. All runs and code - for replicating the results - can be found under https://github.com/MeteSertkan/meta-analysis-based-recsys-eval.

Acknowledgments

This research is supported by the Christian Doppler Research Association (CDG), and has received funding from the EU's H2020 research and innovation program (Grant No. 822670).

References

[1] D. Jannach, C. Bauer, Escaping the mcnamara fallacy: Towards more impactful recommender systems research, AI Magazine 41 (2020) 79–95. URL: https://ojs.aaai.org/aimagazine/index.php/aimagazine/article/view/5312. doi:10.1609/aimag.v41i4.5312.

[2] J. L. Herlocker, J. A. Konstan, L. G. Terveen, J. T. Riedl, Evaluating collaborative filtering recommender systems, ACM Trans. Inf. Syst. 22 (2004) 5–53. URL: https://doi.org/10.1145/963770.963772. doi:10.1145/963770.963772.

[3] E. Zangerle, C. Bauer, Evaluating recommender systems: Survey and framework, ACM Comput. Surv. 55 (2022). URL: https://doi.org/10.1145/3556536. doi:10.1145/3556536.

[4] M. Borenstein, L. V. Hedges, J. P. Higgins, H. R. Rothstein, Introduction to meta-analysis, John Wiley & Sons, Ltd, 2009. URL: https://onlinelibrary.wiley.com/doi/book/10.1002/9780470743386. doi:https://doi.org/10.1002/9780470743386.

[5] M. Sertkan, S. Althammer, S. Hofstätter, Ranger: A toolkit for effect-size based multi-task evaluation, 2023. arXiv:2305.15048.

[6] I. Soboroff, Meta-analysis for retrieval experiments involving multiple test collections, in: Proceedings of the 27th ACM International Conference on Information and Knowledge Management, CIKM '18, Association for Computing Machinery, New York, NY, USA, 2018, p. 713–722. URL: https://doi.org/10.1145/3269206.3271719. doi:10.1145/3269206.3271719.

[7] S. Lewis, M. Clarke, Forest plots: trying to see the wood and the trees, BMJ 322 (2001) 1479–1480. URL: https://www.bmj.com/content/322/7300/1479. doi:10.1136/bmj.322.7300.1479. arXiv:https://www.bmj.com/content/322/7300/1479.full.pdf.

[8] A. Gunawardana, G. Shani, A survey of accuracy evaluation metrics of recommendation tasks, J. Mach. Learn. Res. 10 (2009) 2935–2962.

[9] A. Gunawardana, G. Shani, S. Yogev, Evaluating Recommender Systems, Springer US, New York, NY, 2022, pp. 547–601. URL: https://doi.org/10.1007/978-1-0716-2197-4_15. doi:10.1007/978-1-0716-2197-4_15.

[10] J. Beel, S. Langer, M. Genzmehr, B. Gipp, C. Breitinger, A. Nürnberger, Research paper recommender system evaluation: A quantitative literature survey, in: Proceedings of the International Workshop on Reproducibility and Replication in Recommender Systems Evaluation, RepSys '13, Association for Computing Machinery, New York, NY, USA, 2013, p. 15–22. URL: https://doi.org/10.1145/2532508.2532512. doi:10.1145/2532508.2532512.

[11] J. Beel, B. Gipp, S. Langer, C. Breitinger, Research-paper recommender systems: a literature survey, International Journal on Digital Libraries 17 (2016) 305–338. URL: https://doi.org/10.1007/s00799-015-0156-0. doi:10.1007/s00799-015-0156-0.

[12] L. Michiels, R. Verachtert, B. 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, Association for Computing Machinery, New York, NY, USA, 2022, p. 648–651. URL: https://doi.org/10.1145/3523227.3551472. doi:10.1145/3523227.3551472.

[13] V. W. Anelli, A. Bellogín, A. Ferrara, D. Malitesta, F. A. Merra, C. Pomo, F. M. Donini, T. D. Noia, Elliot: A comprehensive and rigorous framework for reproducible recommender systems evaluation, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2405–2414. URL: https://doi.org/10.1145/3404835.3463245. doi:10.1145/3404835.3463245.

[14] A. Salah, Q.-T. Truong, H. W. Lauw, Cornac: A comparative framework for multimodal recommender systems, Journal of Machine Learning Research 21 (2020) 1–5.

[15] W. X. Zhao, S. Mu, Y. Hou, Z. Lin, Y. Chen, X. Pan, K. Li, Y. Lu, H. Wang, C. Tian, Y. Min, Z. Feng, X. Fan, X. Chen, P. Wang, W. Ji, Y. Li, X. Wang, J.-R. Wen, Recbole: Towards a unified, comprehensive and efficient framework for recommendation algorithms, 2021. arXiv:2011.01731.

[16] L. V. Hedges, Distribution theory for glass's estimator of effect size and related estimators, journal of Educational Statistics 6 (1981) 107–128.

[17] R. DerSimonian, N. Laird, Meta-analysis in clinical trials revisited, Contemporary Clinical Trials 45 (2015) 139–145. URL: https://www.sciencedirect.com/science/article/pii/S1551714415300781. doi:https://doi.org/10.1016/j.cct.2015.09.002, 10th Anniversary Special Issue.

[18] Y. Koren, R. Bell, C. Volinsky, Matrix factorization techniques for recommender systems, Computer 42 (2009) 30–37. doi:10.1109/MC.2009.263.

[19] S. Rendle, C. Freudenthaler, Z. Gantner, L. Schmidt-Thieme, Bpr: Bayesian personalized ranking from implicit feedback, in: Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI '09, AUAI Press, Arlington, Virginia, USA, 2009, p. 452–461.