

Recommender Systems Alone Are Not Everything: Towards a Broader Perspective in the Evaluation of Recommender Systems

Benedikt Loepp¹

¹University of Duisburg-Essen, Duisburg, Germany

Abstract

Thus far, in most of the user experiments conducted in the area of recommender systems, the respective system is considered as an isolated component, i.e., participants can only interact with the recommender that is under investigation. This fails to recognize the situation of users in real-world settings, where the recommender usually represents only one part of a greater system, with many other options for users to find suitable items than using the mechanisms that are part of the recommender, e.g., liking, rating, or critiquing. For example, in current web applications, users can often choose from a wide range of decision aids, from text-based search over faceted filtering to intelligent conversational agents. This variety of methods, which may equally support users in their decision making, raises the question of whether the current practice in recommender evaluation is sufficient to fully capture the user experience. In this position paper, we discuss the need to take a broader perspective in future evaluations of recommender systems, and raise awareness for evaluation methods which we think may help to achieve this goal, but have not yet gained the attention they deserve.

Keywords

Recommender systems, Information filtering, Conversational user interfaces, Decision aids, Evaluation, User experience, User studies, User-centered design

1. Problem statement

Over the last few years, user-centered evaluation of recommender systems has become more and more accepted in the research community [1]. However, it has thus far mostly been ignored, that in real-world settings, recommender systems alone are not everything: Indeed, it is widely accepted that recommendations are responsible for a large amount of the products bought on *Amazon* or the content watched on *Netflix* [2, 3], but there exists a broad range of other methods to help users in making a decision when confronted with an overwhelmingly large item space. These decision aids, however, are studied and developed mostly independently of recommender systems, e.g., text-based search and faceted filtering in the field of information retrieval [4, 5], dialog-based assistants and intelligent chatbots by the conversational user interfaces community [6, 7]. This is mirrored in commercial environments, where one can rarely observe that users

Perspectives on the Evaluation of Recommender Systems Workshop (PERSPECTIVES 2022), September 22nd, 2022, co-located with the 16th ACM Conference on Recommender Systems, Seattle, WA, USA.


✉ benedikt.loepp@uni-due.de (B. Loepp)

🌐 <https://benedikt.loepp.eu/> (B. Loepp)

🆔 0000-0001-9059-5324 (B. Loepp)



© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

are supported in a holistic fashion: Even if multiple decision aids are available, they are often isolated, i.e., the input provided to one component hardly affects the results generated by another. For instance, when a user constrains the set of items by selecting several filter criteria, a recommender that is part of the same system will not necessarily reflect this selection in the generated recommendations, but only take the user’s implicit or explicit item feedback into account (and vice versa). Similarly, the interaction with a chatbot will usually start from nothing, ignoring any other interests or needs that have been expressed before.

This separation is difficult to understand in light of the fact that it is well known that different decision aids contribute differently to the user’s progress in accomplishing typical choice and decision-making tasks [8, 9]. Recent studies on systems that provide multiple support components have confirmed that users use different mechanisms before settling on an item, and that this is (partly) due to personal and situational characteristics [10, 11, 12, 13]. For these reasons, there have, of course, been several calls over the years to bring the methods and their fields closer together [14, 15, 16, 17]. These calls, however, have largely been focused on methodological aspects, whereas it should not be overseen, that this narrow perspective is also reflected in the evaluation of the methods—regardless of whether they are tightly integrated with other decision aids, or appear in a decoupled fashion, as it is currently common.

The work presented in [18] is one of the few exceptions from information retrieval research that argues for a more holistic evaluation approach, in particular, “to think outside the (search) box.” In general, however, this aspect has, to the best of our knowledge, not yet received much attention, neither in the area of information retrieval, nor conversational user interfaces, nor recommender systems [cf. 19, 20, 1, 21]. This means that whenever user experience of recommender systems is studied in empirical experiments, participants can typically only interact with the recommendation component, i.e., the subject of the evaluation, which is often designed specifically for the purpose of the study, especially in academia. Switching to a different method, which participants may find more appealing depending on their personal preferences and progress in the decision-making process, is, however, not possible. Hence, there is a lack of data to indicate which method works best for which user. As a consequence, we argue that it is necessary for *all* kinds of decision aids “to think outside the box.”

In the following, as another motivating example, we present qualitative feedback from a user experiment on multiple decision aids that we recently conducted. Afterwards, we provide an overview of methods that may help to apply a broader perspective when evaluating recommender systems, and thus, to obtain a more accurate picture of real-world scenarios, in which these systems usually represent only one out of many available decision aids.

2. Motivating example and potential future directions

Because of the reasons described above, we argue that a broader perspective is required when evaluating recommender systems. Otherwise, there is the danger of continuing to follow certain paths in the development of interactive and conversational recommendation approaches (cf. the surveys in [22, 23]), without knowing whether these approaches are really what users want. To highlight that this is an issue already, we refer to a recent experiment, in which we asked participants ($n = 100$, 47 females, 2 non-binary, age: $M = 35.36$, $SD = 12.60$) to use different

decision aids. For this purpose, we confronted them with one of two tasks, either related to a goal-driven or an explorative scenario. In both scenarios, participants had to find at least two suitable laptops. To accomplish the respective task, they were allowed to choose (and to switch) between: a faceted filtering component, a content-based recommender with the option to like and dislike items, a product advisor with a dialog showing a limited number of guiding questions, and a natural language chatbot implemented using *Google Dialogflow*.

2.1. Qualitative insights from an experiment with multiple decision aids

While the study is described in more detail in [13], we here want to provide additional comments made by participants when they were asked why they did not use one of the components. For example, this was the case because they were reluctant to use *chatbots*. One participant stated: “I generally do not like interacting with chatbots. I feel whatever input I give to select a product, I might as well use a filter component.” In contrast to the increasing popularity of conversational agents for recommendation purposes, another participant indicated that he or she uses chatbots only when there is “a problem with the product or a payment issue, [but that a] chatbot is not required [for] browsing.” Others were even more direct in their criticism, writing: “I hate chatbots. I feel I spend a lot of time typing, and I hate the fake cheeriness of them,” or that they turn “something really simple like searching for a new laptop into a dark comedy.”

Other participants cited personal characteristics as reasons for their concerns, such as domain knowledge (“I assume the chatbot would be more useful for someone who does not know what to look for.”) or need for control (“I prefer doing things myself, I can chat to the bot when I cannot seem to find what I am looking for using other [...] options.”) For the other options, however, we obtained similar feedback. For example, with respect to the *advisor*, one participant stated that he or she “should have tried this component, but [likes] to shop for these products using filters.” Although it was clearly an academic study, another participant refrained from using the *recommender* because he or she suspected “that this component uses sponsored companies which pay money to have their products included in the recommendation section.” Also in this case, others were generally reluctant, writing: “I never use recommendations as they are never in line with what I need. It may be the bestseller for the store in question, but not for my needs.” Domain knowledge again played a role, notably in both directions, with participants stating to be “knowledgeable enough [that they] do not need recommendations,” but also that they “do not know enough about computers to give a thumbs up or thumbs down.” A lack of knowledge was also a major reason to stay away from *faceted filtering*, of which participants “thought you need technical understanding of laptops,” or mentioned that they “often feel overwhelmed using filters, and generally do not know where to start in regard to buying laptops.”

All these comments suggest that users often have an idea of which decision aid to use. However, this is not necessarily the one that is offered by a system—or is the subject of the experiment they participate in. For a holistic user-centered evaluation, this means, that the perspective is too narrow, as only the user experience with the specific recommender for a given task is addressed, ignoring the interdependencies with other methods that may exist, and may be more suitable depending on the current circumstances. Some participants explicitly indicated that they would like to “specify certain filters and let the recommendations pick only from the result set of the filters,” or to “combine components [so that] the advisor/chatbot works

within the result set created by the must-have filters.” However, this is exactly what participants usually *cannot* do, since most experiments are strongly focused on individual decision aids. To evaluate recommender systems from a broader perspective, we therefore suggest to improve the current practice by applying the following evaluation methods, which, to date, are used only for single methods, or not at all.

2.2. Offline experiments and simulation studies: Richer data, entire systems

Offline experiments are well established in recommender research [24]. However, they are increasingly criticized as they do not allow obtaining insights into the quality dimensions that are relevant from a user perspective [1, 25, 26]. Nevertheless, with the large datasets that are available today (e.g., *MovieLens*, *Netflix*, *Amazon*), they remain essential to make objective decisions whether or not to use a specific recommendation method. However, most datasets are limited to implicit or explicit user-item feedback. Even though they represent different domains, contain a varying amount of side information, and are nowadays often available in sequential form, this limits what can be concluded from the corresponding experiments, i.e., which algorithm generates better *item* recommendations. Therefore, we argue that future offline experiments should be conducted based on richer datasets, which include data from all components of a system, such as both user-item feedback *and* search queries issued. Thus, provided adequate metrics are found, one could also determine which decision aids work best for individual users and keep them longer engaged. More generally, one could examine system support above the item level, e.g., with respect to the objective quality of recommendations for item features, or even of recommendations for switching to other decision aids.

The same applies to *simulation studies*, which have only recently gained more attention in recommender research [27]. By simulating typical user behavior, e.g., with respect to critiquing mechanisms [28] or interaction with items over time [29], this type of experiment has shown strong potential in economic terms. With multiple decision aids, and thus, a larger design space, this will become even more important, also on a global level, e.g., to study long-term user behavior with respect to the question of when a recommender is used, or another component is perceived as more suitable and may contribute more to the user’s progress. In addition, domain and other factors such as product type (search vs. experience), product category (cheap streaming content vs. expensive high-risk products), and the given task (goal-oriented or explorative) may affect which method works best. Thus, simulation studies may be the only way to investigate how preferences evolve over time when using different decision aids, and to understand which interaction effects can occur between a recommender and other components—something that would never be possible with actual users.

2.3. User-centered evaluation: Multiple decision aids, insightful methods

Well-known qualitative methods from human-computer interaction research, which are frequently used at the beginning of user-centered design processes, e.g., *focus groups*, *interviews*, or *contextual inquiries* [cf. 30], are rarely used in the area of recommender systems. We argue, however, that these techniques could be useful for obtaining insights into users’ actual needs with respect to the interaction with such a system, in particular, when it comes to the relation to

other decision aids. In this context, it is worth noting that it has been found only recently, that users' *mental models* do not necessarily correspond to the implementations of recommender systems, and are subject to large inter-individual differences [31]. However, identifying the understanding users have of the system behavior is considered highly important for evaluating the impact of a recommender and improving it [32]. Accordingly, to better inform the design of applications that embed multiple support components, it will be inevitable to explore these models in more depth, in particular, with a focus on the users' comprehension of possible interactions between the components—by means of both qualitative methods such as grounded theory [33] and quantitative approaches such as proposed in [34].

Once a (prototypical) recommender is implemented, *questionnaire-based assessment* is the most common way of measuring the different qualities related to user experience. For this purpose, well-established frameworks and questionnaires exist [e.g., 35, 36, 37], which, however, are strongly focused on dimensions that are specific to recommender systems. On the other hand, general usability questionnaires, e.g., *SUS* [38] and *UEQ* [39], are too broad to draw conclusions about the suitability of this or other decision aids for preferential choice and decision-making tasks. Therefore, existing instruments need to be extended in order to allow for a more global, subjective assessment of recommendation components—in the context of the applications in which they are embedded, and thus, of the interplay with other methods. Otherwise, it will be hardly possible to gain insights into why users prefer a specific method to perform a certain task, and, more generally, whether they would use more strongly connected decision aids, or dislike this idea, e.g., because they expect higher complexity or have increasing privacy concerns.

Finally, we would like to highlight two further aspects that do not receive much attention in current recommender research: *in-situ* and long-term evaluation. The former is important, among others, because questionnaires suffer from the problem that they typically require self reflection disconnected from actual system usage, and, worse, often from consumption or experience of the items, which has shown to influence the assessment of recommendations [40]. Therefore, we deem it necessary to develop methods for a quantitative *in-situ assessment* of the users' motivation to use different components, e.g., via questionnaires directly embedded into the respective applications, as it has been done to study the reasons to switch between search engines [41] or selected decision aids [13]. This, however, may need to be complemented by qualitative methods, such as the *think aloud procedure*, *systematic user observation*, or *shadowing*—techniques that are generally popular but have also rarely been applied in recommender research. Moreover, *eye tracking* may be considered as a useful alternative. Being less disruptive, it has become more popular in recent years, e.g., in studies on recommender interfaces and recommendation presentation [42, 43], critiquing [44, 45], and effects of personal characteristics [46]. However, also in these cases, participants' behavior was observed only in relation to the recommendation component, largely ignoring its surroundings. Either way, following these directions will not be sufficient to understand how users interact with these surroundings over a longer period of time. Therefore, while *repeated study designs*, *longitudinal studies*, and *field studies* are still rare in recommender research, with only very few exceptions [e.g., 47, 48], these methods appear to be of particular importance when taking a broader perspective: Goals and tasks may vary over time, which can have a substantial impact on the usage of different components. Thus, experimental data that represent long-term user behavior only with respect to a *single* decision aid may distort the picture, since other components may be perceived as

more appropriate in other contexts or stages of the decision-making process.

3. Conclusions

Overall, it seems important for future research in the recommender area, but also for other communities, to face the challenge of evaluating the respective methods in a context that is more similar to real-world settings, where decision aids rarely stand on their own. In this position paper, we explained why we think this way, and outlined how this challenge may be addressed. By this means, we hope to raise awareness that contemporary decision aids not only need to be brought together from a methodological perspective, but that a broader perspective is required when evaluating the methods. Of course, this may open up new issues. For instance, the more holistic the evaluation, the higher effort and costs for running an experiment, which are factors that already limit many studies in academia. Moreover, the difficulty of designing an experiment and analyzing its results, but also, the number of application-specific parameters and possible confounding factors, increase with the consideration of more than one decision aid. For these reasons, it will remain important to keep in mind the specific circumstances in which most experiments take place, and not to think that a broader perspective in the evaluation of the methods automatically allows more general conclusions to be drawn.

Nevertheless, we are certain that “thinking outside the (recommendations) box” will help gain a better understanding not only of the degree to which a recommender can satisfy a user in a given situation, but, in particular, of how the interplay with other decision aids can affect the assessment of the system. In the end, this may shift the focus away from further improving decision aids that are less effective or users do not want to use for specific tasks, to those that exhibit the greatest potential for providing support at the respective stage of the decision-making process. For now, however, we hope to encourage at least a discussion about using the mentioned evaluation methods more extensively to gain more in-depth insights into the users’ understanding of and preference for recommendation components in relation to other decision aids. Of course, other methods may equally well be used, but we leave it for future work to provide more concrete suggestions on which methods to use and in which order.

Acknowledgments

Thanks to Timm Kleemann, who implemented the system for the study that is mentioned here, and contributed to this study to the same extent as the author of the present paper. The study was partially supported by the Eurostars project ACODA (grant no. 01QE1946C).

References

- [1] B. P. Knijnenburg, M. C. Willemsen, Evaluating recommender systems with user experiments, in: F. Ricci, L. Rokach, B. Shapira (Eds.), *Recommender Systems Handbook*, Springer US, Boston, MA, USA, 2015, pp. 309–352.

- [2] C. A. Gomez-Urbe, N. Hunt, The Netflix recommender system: Algorithms, business value, and innovation, *ACM Transactions on Management Information Systems* 6 (2015) 13:1–13:19.
- [3] B. Smith, G. Linden, Two decades of recommender systems at Amazon.com, *IEEE Internet Computing* 21 (2017) 12–18.
- [4] R. Baeza-Yates, B. Ribeiro-Neto, *Modern Information Retrieval*, ACM, New York, NY, USA, 1999.
- [5] M. A. Hearst, *Search User Interfaces*, Cambridge University Press, Cambridge, UK, 2009.
- [6] K. Ramesh, S. Ravishankaran, A. Joshi, K. Chandrasekaran, A survey of design techniques for conversational agents, in: *ICICCT '17: Proceedings of the 2nd International Conference on Information, Communication and Computing Technology*, Springer Singapore, Singapore, 2017, pp. 336–350.
- [7] A. Anand, L. Cavedon, H. Joho, M. Sanderson, B. Stein, Conversational search (Dagstuhl Seminar 19461), *Dagstuhl Reports* 9 (2020) 34–83.
- [8] G. Häubl, V. Trifts, Consumer decision making in online shopping environments: The effects of interactive decision aids, *Marketing Science* 19 (2000) 4–21.
- [9] S. Castagnos, N. Jones, P. Pu, Recommenders' influence on buyers' decision process, in: *RecSys '09: Proceedings of the 3rd ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2009, pp. 361–364.
- [10] J. Schaffer, J. Humann, J. O'Donovan, T. Höllerer, Quantitative modeling of dynamic human-agent cognition, in: M. D. McNeese, E. Salas, M. R. Endsley (Eds.), *Contemporary Research: Models, Methodologies, and Measures in Distributed Team Cognition*, CRC Press, Boca Raton, FL, USA, 2020, pp. 137–186.
- [11] P. Virdi, A. D. Kalro, D. Sharma, Online decision aids: The role of decision-making styles and decision-making stages, *International Journal of Retail & Distribution Management* 48 (2020) 555–574.
- [12] T. Kleemann, M. Wagner, B. Loepp, J. Ziegler, Modeling user interaction at the convergence of filtering mechanisms, recommender algorithms and advisory components, in: *Mensch & Computer 2021 – Tagungsband*, ACM, New York, NY, USA, 2021, pp. 531–543.
- [13] T. Kleemann, B. Loepp, J. Ziegler, Towards multi-method support for product search and recommending, in: *UMAP '22: Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY, USA, 2022, pp. 74–79.
- [14] H. Garcia-Molina, G. Koutrika, A. Parameswaran, Information seeking: Convergence of search, recommendations, and advertising, *Communications of the ACM* 54 (2011) 121–130.
- [15] E. H. Chi, Blurring of the boundary between interactive search and recommendation, in: *IUI '15: Proceedings of the 20th International Conference on Intelligent User Interfaces*, ACM, New York, NY, USA, 2015, p. 2.
- [16] B. Loepp, On the convergence of intelligent decision aids, in: *UCAI '21: Proceedings of the 2nd Workshop on User-Centered Artificial Intelligence*, 2021.
- [17] A. D. Starke, M. Lee, Unifying recommender systems and conversational user interfaces, in: *CUI '22: Proceedings of the 4th International Conference on Conversational User Interfaces*, ACM, New York, NY, USA, 2022.
- [18] P. Clough, Evaluation: Thinking outside the (search) box, in: *FIRE '14: Proceedings of the*

- Forum for Information Retrieval Evaluation, ACM, New York, NY, USA, 2015, pp. 1–9.
- [19] D. Kelly, Methods for evaluating interactive information retrieval systems with users, *Foundations and Trends in Information Retrieval* 3 (2009) 1–224.
 - [20] C. Mulwa, S. Lawless, M. Sharp, V. Wade, The evaluation of adaptive and personalised information retrieval systems: A review, *International Journal of Knowledge and Web Intelligence* 2 (2011) 138–156.
 - [21] A. B. Kocaballi, L. Laranjo, E. Coiera, Understanding and measuring user experience in conversational interfaces, *Interacting with Computers* 31 (2019) 192–207.
 - [22] M. Jugovac, D. Jannach, Interacting with recommenders – Overview and research directions, *ACM Transactions on Interactive Intelligent Systems* 7 (2017) 10:1–10:46.
 - [23] D. Jannach, A. Manzoor, W. Cai, L. Chen, A survey on conversational recommender systems, *ACM Computing Surveys* 54 (2022) 105:1–105:36.
 - [24] A. Gunawardana, G. Shani, S. Yogev, Evaluating recommender systems, in: F. Ricci, L. Rokach, B. Shapira (Eds.), *Recommender Systems Handbook*, Springer, New York, NY, USA, 2022, pp. 547–601.
 - [25] M. Rossetti, F. Stella, M. Zanker, Contrasting offline and online results when evaluating recommendation algorithms, in: *RecSys ’16: Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2016, pp. 31–34.
 - [26] T. Rehorek, O. Biza, R. Bartyzal, P. Kordik, I. Povalyev, O. Podstavek, Comparing offline and online evaluation results of recommender systems, in: *REVEAL ’18: Proceedings of the Workshop on Offline Evaluation for Recommender Systems*, 2018.
 - [27] N. Hazrati, F. Ricci, Simulating users’ interactions with recommender systems, in: *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY, USA, 2022, pp. 95–98.
 - [28] H. Xie, D. D. Wang, Y. Rao, T.-L. Wong, L. Y. K. Raymond, L. Chen, F. L. Wang, Incorporating user experience into critiquing-based recommender systems: A collaborative approach based on compound critiquing, *International Journal of Machine Learning and Cybernetics* 9 (2018) 837–852.
 - [29] J. McInerney, E. Elahi, J. Basilico, Y. Raimond, T. Jebara, Accordion: A trainable simulator for long-term interactive systems, in: *RecSys ’21: Proceedings of the 15th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2021, pp. 102–113.
 - [30] G. Cockton, Usability evaluation, *The Encyclopedia of Human-Computer Interaction* (2nd Edition) (2013).
 - [31] T. Ngo, J. Kunkel, J. Ziegler, Exploring mental models for transparent and controllable recommender systems: A qualitative study, in: *UMAP ’20: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY, USA, 2020, pp. 183–191.
 - [32] M. M. Ghorri, A. Dehpanah, J. Gemmell, H. Qahri-Saremi, B. Mobasher, Does the user have a theory of the recommender? A grounded theory study, in: *Adjunct Proceedings of the 30th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY, USA, 2022, pp. 167–174.
 - [33] J. Corbin, A. Strauss, *Basics of Qualitative Research: Techniques and Procedures for Developing Grounded Theory*, 3 ed., Sage Publications, Inc., Thousand Oaks, CA, USA, 2008.

- [34] J. Kunkel, T. Ngo, J. Ziegler, N. Krämer, Identifying group-specific mental models of recommender systems: A novel quantitative approach, in: C. Ardito, R. Lanzilotti, A. Malizia, H. Petrie, A. Piccinno, G. Desolda, K. Inkpen (Eds.), *Human-Computer Interaction – INTERACT 2021*, volume 12935 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2021, pp. 383–404.
- [35] B. P. Knijnenburg, M. C. Willemsen, A. Kobsa, A pragmatic procedure to support the user-centric evaluation of recommender systems, in: *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2011, pp. 321–324.
- [36] P. Pu, L. Chen, R. Hu, A user-centric evaluation framework for recommender systems, in: *RecSys '11: Proceedings of the 5th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2011, pp. 157–164.
- [37] B. P. Knijnenburg, M. C. Willemsen, Z. Gantner, H. Soncu, C. Newell, Explaining the user experience of recommender systems, *User Modeling and User-Adapted Interaction* 22 (2012) 441–504.
- [38] J. Brooke, SUS – A quick and dirty usability scale, in: *Usability Evaluation in Industry*, Taylor & Francis, London, UK, 1996, pp. 189–194.
- [39] B. Laugwitz, T. Held, M. Schrepp, Construction and evaluation of a user experience questionnaire, in: A. Holzinger (Ed.), *HCI and Usability for Education and Work*, volume 5298 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2008, pp. 63–76.
- [40] B. Loepp, T. Donkers, T. Kleemann, J. Ziegler, Impact of item consumption on assessment of recommendations in user studies, in: *RecSys '18: Proceedings of the 12th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2018, pp. 49–53.
- [41] Q. Guo, R. W. White, Y. Zhang, B. Anderson, S. T. Dumais, Why searchers switch: Understanding and predicting engine switching rationales, in: *SIGIR '11: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, New York, NY, USA, 2011, pp. 335–344.
- [42] Q. Zhao, S. Chang, F. M. Harper, J. A. Konstan, Gaze prediction for recommender systems, in: *RecSys '16: Proceedings of the 10th ACM Conference on Recommender Systems*, ACM, New York, NY, USA, 2016, pp. 131–138.
- [43] P. Gaspar, M. Kompan, J. Simko, M. Bielikova, Analysis of user behavior in interfaces with recommended items – An eye-tracking study, in: *IntRS '18: Proceedings of the 5th Joint Workshop on Interfaces and Human Decision Making for Recommender Systems*, 2018, pp. 32–36.
- [44] L. Chen, F. Wang, An eye-tracking study: Implication to implicit critiquing feedback elicitation in recommender systems, in: *UMAP '16: Proceedings of the 24th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY, USA, 2016, pp. 163–167.
- [45] L. Chen, F. Wang, W. Wu, Inferring users' critiquing feedback on recommendations from eye movements, in: A. Goel, M. B. Díaz-Agudo, T. Roth-Berghofer (Eds.), *Case-Based Reasoning Research and Development*, volume 9969 of *Lecture Notes in Computer Science*, Springer, Berlin, Germany, 2016, pp. 62–76.
- [46] M. Millecamp, N. N. Htun, C. Conati, K. Verbert, What's in a user? Towards personalising transparency for music recommender interfaces, in: *UMAP '20: Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, ACM, New York, NY,

USA, 2020, pp. 173–182.

- [47] Y. Zhong, T. L. S. Menezes, V. Kumar, Q. Zhao, F. M. Harper, A field study of related video recommendations: Newest, most similar, or most relevant?, in: Proceedings of the 12th ACM Conference on Recommender Systems, ACM, New York, NY, USA, 2018, pp. 274–278.
- [48] Y. Liang, M. C. Willemsen, Exploring the longitudinal effects of nudging on users’ music genre exploration behavior and listening preferences, in: RecSys ’22: Proceedings of the 16th ACM Conference on Recommender Systems, ACM, New York, NY, USA, to appear.