Understanding User Interactions with Podcast Recommendations Delivered Via Voice

Longqi Yang
Cornell Tech, Cornell University
ly283@cornell.edu

Michael Sobolev
Cornell Tech, Cornell University
michael.sobolev@cornell.edu

Christina Tsangouri
City University of New York
christinatsangouri@gmail.com

Deborah Estrin
Cornell Tech, Cornell University
destrin@cornell.edu

ABSTRACT

Voice interfaces introduced by smart speakers present new opportunities and challenges for podcast content recommendations. Understanding how users interact with voice-based recommendations has the potential to inform better design of vocal recommenders. However, existing knowledge about user behavior is mostly for visual interfaces, such as the web, and is not directly transferable to voice interfaces, which rely on user listening and do not support skimming and browsing. To fill in the gap, we conducted a controlled study to compare user interactions with recommendations delivered visually to those with recommendations delivered vocally. Through an online A/B testing with 100 participants, we found that when recommendations are vocally conveyed, users consume more slowly, explore less, and choose fewer long-tail items. The study also reveals the correlation between user choices and exploration via voice interfaces. Our findings pose challenges to the design of voice interfaces, such as adaptively recommending diverse content and designing better navigation mechanisms.

KEYWORDS

Recommendation; Voice interface; User interaction; User behavior

ACM Reference Format:
Longqi Yang, Michael Sobolev, Christina Tsangouri, and Deborah Estrin. 2018. Understanding User Interactions with Podcast Recommendations Delivered Via Voice. In *Twelfth ACM Conference on Recommender Systems (RecSys '18), October 2–7, 2018, Vancouver, BC, Canada.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3240323.3240389

1 INTRODUCTION

Virtual assistants and smart speakers, such as Apple Siri, Amazon Alexa, and Google Home, are becoming increasingly and widely adopted every year. A recent survey estimates that 16 percent of Americans (around 40 million) own a smart speaker and 65 percent of them would not go back to a life without one.[1] These devices introduce a **voice interface** for consuming spoken word content (e.g., podcast and audiobook) when users have limited visual attention. To personalize the user experience, a voice interface usually delivers recommendations through audio using text-to-speech technology. For example, many commercial applications (e.g., Stitcher [4] and AnyPod [3]) present recommendations in a list (Fig. 1): a smart speaker reads out items sequentially, and then a user selects a piece of content explicitly (e.g., *play the evolution of artificial intelligence*) or implicitly (e.g., the speaker plays an episode until the user chooses *next* or *previous*).

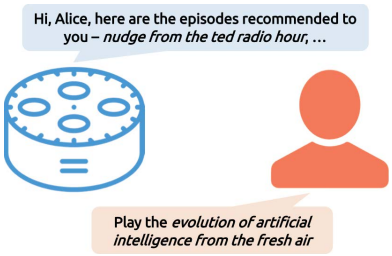

Figure 1: An example scenario where recommendations are delivered via voice. A smart speaker presents a list of recommendations vocally. Then a user selects an item by issuing a command such as *play* or *next*.

Prior research revealed important user interaction patterns with *recommendations on visual interfaces* (**VISUAL**). For example, position bias [10, 22, 24], rating conformity [15, 23], and user exploration and exploitation [18, 20, 21]. These findings have significantly informed the design and evaluation of visual recommendation systems, for example, unbiased evaluation [10, 15, 23] and improved diversity [18, 20, 21]. However, because of the unique characteristics of a voice interface (e.g., it relies on listening, has a narrow information channel, and doesn't allow skimming and browsing), prior research is insufficient to understand user interactions with *recommendations communicated via voice* (**VOICE**). Additional research is needed to address how users are going to consume VOICE and what improvements can be implemented for the current voice-based recommendation delivery paradigm.

In this paper, we conducted a controlled user study to understand user interactions with VOICE as compared to VISUAL. We focused on podcast content because of its increasing importance as a major channel for information and entertainment. For example, there

[1] https://www.nationalpublicmedia.com/smart-audio-report/latest-report/)

RecSys '18, October 2–7, 2018, Vancouver, BC, Canada

ACM ISBN 978-1-4503-5901-6/18/10…$15.00
https://doi.org/10.1145/3240323.3240389

are 67 million monthly and 42 million weekly podcast listeners in the United States [1]. Specifically, we tested three hypotheses (**H**) and investigated two research questions (**Q**) on the evaluation and choices of the content.

- **H1.** VISUAL is more efficient to consume than VOICE.
- **H2.** Users explore fewer items with VOICE than with VISUAL.
- **H3.** Users choose fewer lower-ranked items with VOICE than with VISUAL.
- **Q1.** How many items does a user consider with VOICE before making a choice?
- **Q2.** Does the recommendation medium affect users' satisfaction with choices and the recommendation engine?

Through a between-subject study with 100 participants recruited from the Amazon Mechanical Turk, our main findings include the following: (1) users consumed VISUAL **9 times** as quickly as VOICE, (2) users explored at least **3 times** as much VISUAL as VOICE before making a choice, (3) users chose items **6 times** as deep into VISUAL as they did into VOICE, (4) users' range of exploration on VOICE was **2 times** higher than their actual choice, and (5) there is no evidence to suggest that the consuming medium affected users' satisfaction with their choices or the recommendations. The findings have important implications for the design of voice-based recommendations.

2 RELATED WORK

Our work is inspired by previous research that studied user behavior in consuming visual recommendations. Prior work used eye tracking [6, 9, 24] and log analysis [14, 22, 23] to understand how different visual presentations may affect users' reactions to recommendations. For instance, users' ratings tend to agree or contrast with those already given by other reviewers [15, 23], and the positions of the items are likely to affect users' attention and choices [14, 22]. However, such knowledge of user interactions may not apply to a voice interface, which is fundamentally different from a visual interface. Our work fills in the gap by comparing side-by-side user behavior under the two types of interfaces and discovering unique patterns for VOICE.

Another related direction of research is conversational search and recommendation, which is mainly focused on **learning users' preference and intentions** from spoken queries and conversations [7, 13, 17, 19]. For example, Thompson et al. [19] developed the Adaptive Place Advisor, which assists users in choosing preferable destinations, Christakopoulou et al. [7] proposed a bandit-based algorithm to elicit user preferences toward restaurants, and Kang et al. [13] explored the initial and follow-up queries users tend to issue to a voice agent. Although prior work leveraged an audio medium for preference learning, it did not address the question of how users are going to interact with the recommendations *after the ranked list is finalized and delivered*. Our study complements the existing research by investigating user interaction patterns with recommended items.

3 STUDY DESIGN AND PROCEDURE

We designed a web application that presents a fixed list of podcast recommendations through a visual or voice interface. With the interface as the only controlled variable, we observed and compared how users interact with recommendations. To eliminate external confounding factors, the application interface for either medium was minimally designed and not optimized. This limits the current study in generalizing to state-of-the-art visual designs but provides a fair environment for intrinsic side-by-side comparison. In this section, we present the details of the study procedure, the participant recruitment, and the mechanisms we used to control the quality of the experiment.

Procedure. The study follows a between-subject design. Participants were randomly divided into two groups (A and B) and were instructed to use a web application to *select an episode that you like most and commit to listen for a full 5 min* from a list of recommended options. We used a commitment mechanism to encourage choices according to users' actual preferences. The podcast recommendations were fixed for all participants and were generated by taking the most popular episodes from the top-ranked 152 shows that were available on iTunes on 02/28/17.

Specifically, participants from **Group A** were provided with a visual interface for browsing podcast recommendations (Fig. 2, step 1): The titles of the recommended episodes were presented in a list and were auto-loaded when scrolling. A participant could indicate their decision by typing in corresponding index number. Participants from **Group B** received the same list of recommendations as Group A but through an audio channel (Fig. 2, step 1). To mimic the interactions users may have with smart speakers, we used the Amazon Polly service [2] to convert episode titles into speech. The input text was derived by prefixing the titles of the episodes with positive integers in ascending order, that is, "number 1 [episode 1 title], number 2 [episode 2 title], ..." Participants continuously listened to the generated audio and made a selection using the same approach as Group A. Our design simplified the user responses (in reality, users may respond by speech, as shown in Fig. 1), which made it feasible for the study to be conducted on the web. Such a simplification is not likely to affect the outcomes, since our study was focused on the user interactions up to the time of users' decision.

Participants were required to listen to their selected episode for 5 minutes (navigation to the next portion of the study was blocked until the podcast had been playing for a minimum of 5 minutes, as shown in Fig. 2, step 2) before answering the final evaluation questions, including *how much you liked your podcast choice (on a Likert scale from 1 to 5)*, *why you liked or disliked your podcast choice*, *how much you liked our recommendations (on a Likert scale from 1 to 5)*, and *why you liked or disliked our recommendations* (Fig. 2, step 3). These subjective questions were designed to capture users' general preference. In the player interface (Fig. 2, step 2), participants also had the option of downloading the episode. The study procedure was reviewed and approved by the Institutional Review Board.

Participants. 108 participants were recruited from the Amazon Mechanical Turk platform. To be qualified for the study, participants were required to be from the United States and have an acceptance rate above 95. To further control the quality, we implemented an attention check mechanism: At the end of the study, each participant was required to answer the question, *Please describe the podcast you listened to*. Only those who passed the verification were included in the final analysis. Eventually, 100 (50 in Group A and 50 in Group B) valid participants were verified.

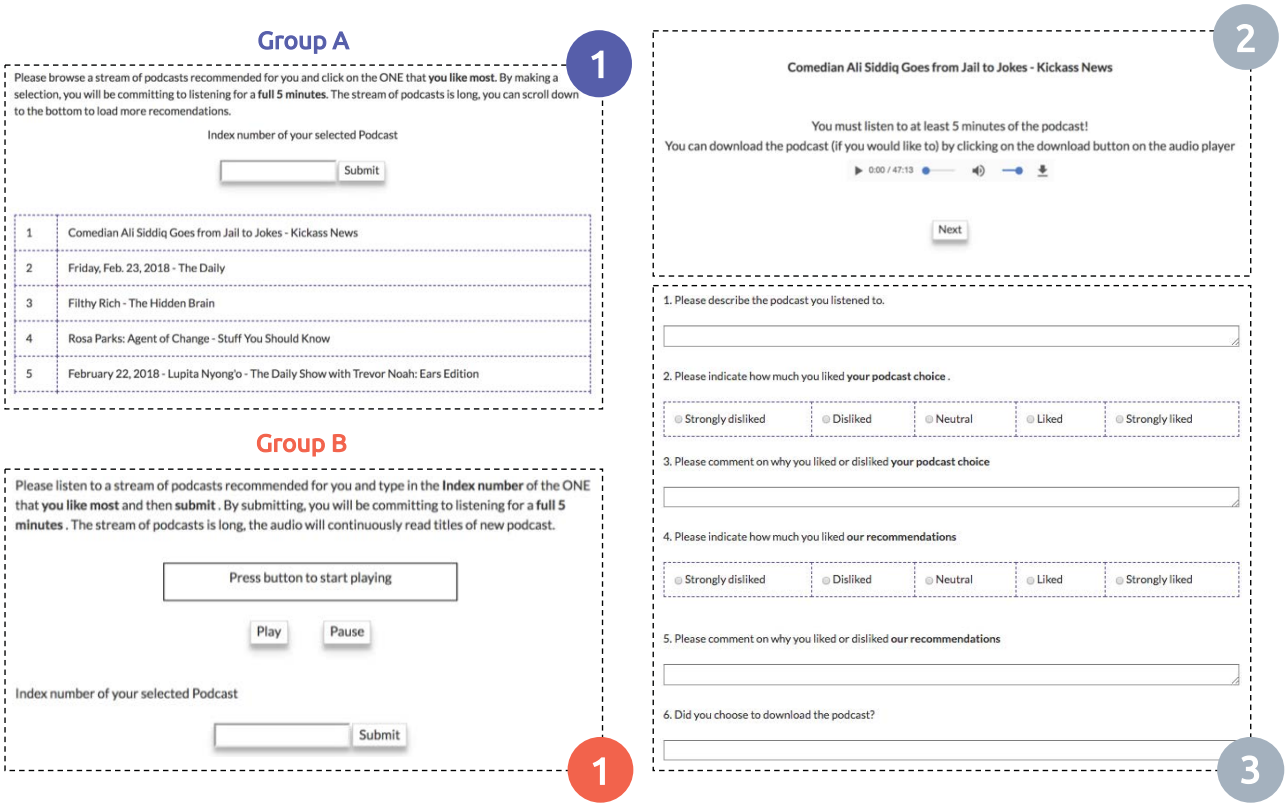

Figure 2: Web interfaces used in the A/B testing. Participants from either group went through three phases: (1) browse (Group A) or listen to (Group B) the titles of the recommended episodes and choose a preferred episode, (2) listen to the chosen episode for at least 5 minutes, and (3) express their level of satisfaction with their choice and the recommendation list.

Table 1: User interaction quantification. Time, Choice ID, Search ID, and Efficiency were measured for VOICE and VISUAL. (Search ID under VISUAL was approximated by Choice ID as a lower bound.)

Medium	Time(s)	Choice ID	Search ID	Efficiency
VOICE	144.8±17.8	7.3±1.4	15.3±1.8	0.1±0.0
VISUAL	77.9±10.0	47.8±6.5	N/A	0.9±0.1

4 EXPERIMENTAL RESULTS

Throughout the experiment, participants' dwelling time, choices, and inputs were recorded. To test our hypotheses and answer our research questions, we measured the following variables:

- **Choice ID:** the index number of the chosen episode.
- **Search ID:** the maximum index number of the recommendations that a participant considers.[2]
- **Time:** the amount of time that a participant spent before making a choice.
- **Satisfaction:** participants' Likert-scale ratings for their choices and the recommendation engine.

[2]Note that for VISUAL, our web application did not accurately capture the Search ID. Therefore, we used Choice ID as a *lower-bound estimate*. Future work could address this limitation by leveraging more sophisticated tracking methods.

For each interface, we averaged these measures over the participants; the results are given in Table 1. Additionally, we illustrate the relationship between Choice ID and Search ID under VOICE in Fig. 3 and the satisfaction distribution in Fig. 4. Next, we discuss our findings regarding the hypotheses and the research questions.

H1. System Efficiency. Efficiency has been a standard metric for the system usability test and is usually measured using the task completion time [8]. In our experimental setting, we measured the efficiency as the speed with which a user explored the recommendations, that is, the average number of recommendations a participant explored per unit time: Efficiency $= \frac{\text{Search ID}}{\text{Time}}$. We hypothesized that VOICE is less efficient to consume than VISUAL because browsing is usually faster than listening; also, that VISUAL allows richer interactions, such as skim and skip. The average efficiency for each interface is given in Table 1. The results demonstrate that, on average, VISUAL users consumed 9 times as many items per unit time as VOICE users — a result that is statistically significant ($p < 10^{-7}$, effect size $d = 1.17$). In other words, VOICE was significantly less efficient than VISUAL. To further understand how such an inefficient interface may affect user choices, we consider **H2** and **H3**.

H2. User Exploration. The first hypothesis we had in regard to user choice behavior on consumption of VOICE is that users

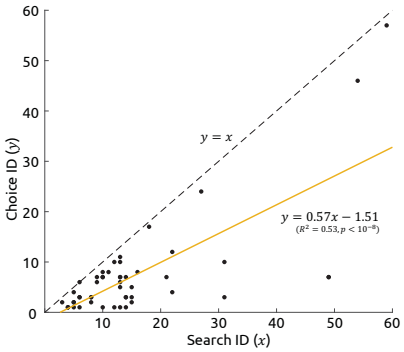

Figure 3: Scatter plot of Choice ID vs. Search ID under VOICE.

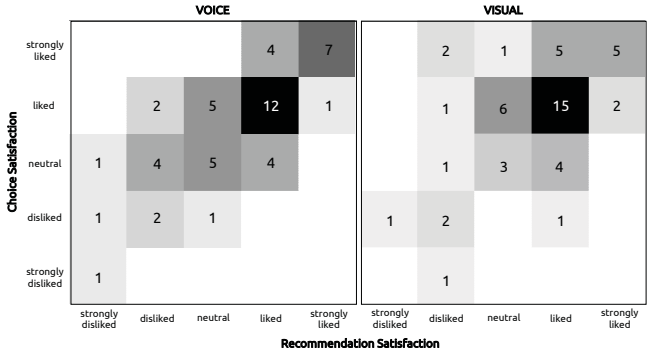

Figure 4: Distribution of the satisfaction ratings for the choice made and the recommendations. Each cell represents the number of participants.

may explore less before making a decision. Encouraging users to broadly explore is a critical step to permeating the potential filter bubbles [16]. As found in previous studies [12, 25], a good recommendation engine should be not only accurate, but also diverse. We used the Search ID as an indicator of the scope of exploration. As shown in Table 1, the average Search ID on VISUAL (where Choice ID was used as the lower-bound) is 2 times larger than the average on VOICE — another result that is statistically significant ($p < 10^{-5}$, effect size $d = 0.95$). Therefore, VOICE significantly hindered wide exploration of the recommendations.

H3. Choice of Lower-ranked Items. We also investigated the actual choices users made, and we hypothesized that users may be less likely to choose lower-ranked items with VOICE. We leveraged the Choice ID to quantify users' choices. According to Table 1, the average Choice ID on VISUAL was 5 to 6 times larger than the average on VOICE. In other words, user consumption was more likely to be concentrated on the top-ranked items under VOICE. Also, since popular items are often ranked higher [5], VOICE consumption may be more biased toward popularity.

Q1. Search Behavior and Choice. Based on the findings in regard to **H1** and **H2**, we further investigated the relationship between user exploration and choice. Specifically, we quantified the size of the users' decision space under VOICE, that is, the number of options a user tended to consider, using the difference and the correlation between the Search ID and the Choice ID. As shown in Table 1, on average, users tended to consider 8 episodes before making their decision. Also, according to the results of the linear regression ($R^2 > 0.5, p < 10^{-8}$) in Fig. 3, the slope is significantly less than 1 ($p < 10^{-8}$). Consistent with the exploitation–exploration trade-off [20], we demonstrated that users did not stop the search process and explored more options before choosing their favorite.

Q2. User satisfaction. Lastly, we examined users' overall satisfaction with the recommendations and their choices. The average ratings are presented in Table 1. We quantitatively measured the satisfaction difference by first transforming each choice on the Likert scales (1–5) to a value between 0 and 1 using the average cumulative proportion [11], and then comparing the average transformed value. The satisfaction with the choices made and the recommendations are not statistically different for VOICE and VISUAL (choice: $p = 0.33$, recommendation: $p = 0.53$). However, this result may differ under more complex and modern visual interfaces, such as, multimodal presentations. Additionally, we discovered that pairs of satisfaction scores have a higher correlation under VOICE ($r = 0.55$, $p < 10^{-9}$) than under VISUAL ($r = 0.20$, $p = 0.001$). Although this finding was not hypothesized, it provides insights into the interaction of the subjective evaluation of the recommendations compared to the content of choices under different interfaces. We plan to explore this in future work.

5 CONCLUSIONS AND DISCUSSIONS

Our user study revealed unique interaction patterns of users consuming recommendations delivered via voice, which suggest design implications for future voice-based recommendations.

- **Adaptive and diverse recommendations.** As shown in the study, under a voice interface, users tend to explore less and to choose higher-ranked items. To encourage exploration and permeate the potential filter bubbles, the top-ranked items should be diverse, personalized, and adaptively changing, so that users are exposed to broad options.
- **Navigation mechanism.** Inefficiency is a critical weak point of current voice-based recommendations. Future interface design should enable users to navigate the recommendation space more easily. For example, organizing presented items with a hierarchical menu structure may improve the fluency.

Our initial study does not address several complex interaction scenarios in the real world, which could be studied in future work: (1) use of a combination of VISUAL and VOICE, such as using Siri on an IPhone, (2) consumption of VOICE under hands-free conditions, such as driving, and (3) use of adaptive updating of recommendations.

ACKNOWLEDGMENTS

We thank the anonymous reviewers for their insightful comments and suggestions. This research was funded by the National Science Foundation (#1700832) and Oath (the Connected Experiences Laboratory at Cornell Tech). The work was further supported by the small data lab at Cornell Tech, which receives funding from NSF, NIH, RWJF, UnitedHealth Group, Google, and Adobe.

REFERENCES

[1] 2017. *Edison Research: The Podcast Consumer 2017.* http://www.edisonresearch.com/the-podcast-consumer-2017/

[2] 2018. *Amazon Polly.* https://aws.amazon.com/polly/

[3] 2018. *AnyPod.* http://anypod.net/manual

[4] 2018. *Stitcher.* https://www.stitcher.com/

[5] Himan Abdollahpouri, Robin Burke, and Bamshad Mobasher. 2017. Controlling Popularity Bias in Learning-to-Rank Recommendation. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 42–46.

[6] Sylvain Castagnos, Nicolas Jones, and Pearl Pu. 2010. Eye-tracking product recommenders' usage. In *Proceedings of the fourth ACM conference on Recommender systems*. ACM, 29–36.

[7] Konstantina Christakopoulou, Filip Radlinski, and Katja Hofmann. 2016. Towards conversational recommender systems. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 815–824.

[8] Erik Frøkjær, Morten Hertzum, and Kasper Hornbæk. 2000. Measuring usability: are effectiveness, efficiency, and satisfaction really correlated?. In *Proceedings of the SIGCHI conference on Human Factors in Computing Systems*. ACM, 345–352.

[9] Laura A Granka, Thorsten Joachims, and Geri Gay. 2004. Eye-tracking analysis of user behavior in WWW search. In *Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 478–479.

[10] Katja Hofmann, Anne Schuth, Alejandro Bellogin, and Maarten De Rijke. 2014. Effects of position bias on click-based recommender evaluation. In *European Conference on Information Retrieval*. Springer, 624–630.

[11] Cheng-Kang Hsieh, Longqi Yang, Honghao Wei, Mor Naaman, and Deborah Estrin. 2016. Immersive recommendation: News and event recommendations using personal digital traces. In *Proceedings of the 25th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 51–62.

[12] Dietmar Jannach, Paul Resnick, Alexander Tuzhilin, and Markus Zanker. 2016. Recommender systemsâĂŤbeyond matrix completion. *Commun. ACM* 59, 11 (2016), 94–102.

[13] Jie Kang, Kyle Condiff, Shuo Chang, Joseph A Konstan, Loren Terveen, and F Maxwell Harper. 2017. Understanding How People Use Natural Language to Ask for Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 229–237.

[14] Kristina Lerman and Tad Hogg. 2014. Leveraging position bias to improve peer recommendation. *PloS one* 9, 6 (2014), e98914.

[15] Yiming Liu, Xuezhi Cao, and Yong Yu. 2016. Are You Influenced by Others When Rating?: Improve Rating Prediction by Conformity Modeling. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 269–272.

[16] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[17] Barry Smyth, Lorraine McGinty, James Reilly, and Kevin McCarthy. 2004. Compound critiques for conversational recommender systems. In *Proceedings of the 2004 IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE Computer Society, 145–151.

[18] Choon Hui Teo, Houssam Nassif, Daniel Hill, Sriram Srinivasan, Mitchell Goodman, Vijai Mohan, and SVN Vishwanathan. 2016. Adaptive, personalized diversity for visual discovery. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 35–38.

[19] Cynthia A Thompson, Mehmet H Goker, and Pat Langley. 2004. A personalized system for conversational recommendations. *Journal of Artificial Intelligence Research* 21 (2004), 393–428.

[20] Longqi Yang, Yin Cui, Fan Zhang, John P Pollak, Serge Belongie, and Deborah Estrin. 2015. Plateclick: Bootstrapping food preferences through an adaptive visual interface. In *Proceedings of the 24th acm international on conference on information and knowledge management*. ACM, 183–192.

[21] Longqi Yang, Cheng-Kang Hsieh, Hongjian Yang, John P Pollak, Nicola Dell, Serge Belongie, Curtis Cole, and Deborah Estrin. 2017. Yum-me: a personalized nutrient-based meal recommender system. *ACM Transactions on Information Systems (TOIS)* 36, 1 (2017), 7.

[22] Yisong Yue, Rajan Patel, and Hein Roehrig. 2010. Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In *Proceedings of the 19th international conference on World wide web*. ACM, 1011–1018.

[23] Xiaoying Zhang, Junzhou Zhao, and John Lui. 2017. Modeling the Assimilation-Contrast Effects in Online Product Rating Systems: Debiasing and Recommendations. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*. ACM, 98–106.

[24] Qian Zhao, Shuo Chang, F Maxwell Harper, and Joseph A Konstan. 2016. Gaze prediction for recommender systems. In *Proceedings of the 10th ACM Conference on Recommender Systems*. ACM, 131–138.

[25] Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences* 107, 10 (2010), 4511–4515.