# Effects of Personalized and Aggregate Top-N Recommendation Lists on User Preference Ratings

GEDIMINAS ADOMAVICIUS, University of Minnesota, Minneapolis, Minnesota
JESSE BOCKSTEDT, Emory University, Atlanta, Georgia
SHAWN CURLEY, University of Minnesota, Minneapolis, Minnesota
JINGJING ZHANG, Indiana University, Bloomington, Indiana



Prior research has shown a robust effect of personalized product recommendations on user preference judgments for items. Specifically, the display of system-predicted preference ratings as item recommendations has been shown in multiple studies to bias users' preference ratings after item consumption in the direction of the predicted rating. Top-N lists represent another common approach for presenting item recommendations in recommender systems. Through three controlled laboratory experiments, we show that top-N lists do not induce a discernible bias in user preference judgments. This result is robust, holding for both lists of personalized item recommendations and lists of items that are top-rated based on averages of aggregate user ratings. Adding numerical ratings to the list items does generate a bias, consistent with earlier studies. Thus, in contexts where preference biases are of concern to an online retailer or platform, top-N lists, without numerical predicted ratings, would be a promising format for displaying item recommendations.

CCS Concepts: • **Information systems → Information retrieval**; **Retrieval tasks and goals**; **Recommender systems**; **Evaluation of retrieval results**; **Presentation of retrieval results**;

Additional Key Words and Phrases: Recommender systems, top-N recommendations, user preferences, personalization, decision biases

**ACM Reference format:**
Gediminas Adomavicius, Jesse Bockstedt, Shawn Curley, and Jingjing Zhang. 2021. Effects of Personalized and Aggregate Top-N Recommendation Lists on User Preference Ratings. *ACM Trans. Inf. Syst.* 39, 2, Article 13 (January 2021), 38 pages.
https://doi.org/10.1145/3430028

## 1 INTRODUCTION

Recent studies show that interacting with online personalization and recommendation systems can have unintended effects on user preferences and economic behavior [Adomavicius et al. 2013, 2018, 2019; Chen et al. 2013; Cosley et al. 2003; Guo and Dunson 2015; Jannach et al. 2015; Schnabel et al. 2016]. In particular, users' self-reported judgments can be significantly distorted by the

Authors' addresses: G. Adomavicius and S. Curley, Department of Information and Decision Sciences, Carlson School of Management, University of Minnesota, 321 19th Avenue South, Minneapolis, MN 55455; emails: {gedas; curley}@umn.edu; J. Bockstedt, Information Systems and Operations Management, Goizueta Business School, Emory University, 1300 Clifton Rd, Atlanta, GA 30322; email: bockstedt@emory.edu; J. Zhang, Department of Operations and Decision Technologies, Kelley School of Business, Indiana University, 1309 E. 10th Street, Bloomington, IN 47405; email: jjzhang@indiana.edu.

1046-8188/2021/01-ART13 $15.00
https://doi.org/10.1145/3430028

previously observed system's predictions of users' preference ratings. For example, Adomavicius et al. [2013] found evidence that a recommendation provided by an online system affects users' ratings for products, even immediately following consumption. Additionally, Adomavicius et al. [2018] found that personalized ratings displayed to users significantly swayed their willingness to pay for digital songs in the direction of the system-displayed rating value, even after sampling the song (i.e., listening to a fragment of the song) prior to pricing. This post-consumption effect of ratings is particularly problematic. Certainly, it is desirable for system recommendations to impact user behavior prior to consumption—helping the users choose the most relevant items to consume is the main intended purpose of recommender systems. However, after consumption (especially *immediately* after consumption, when the uncertainty about the item is arguably the lowest) this effect is not desirable and constitutes a bias in users' expressed preferences. This effect also goes against the widely accepted assumption that the user-specified ratings represent reliable "ground truth" about users' actual preferences, as these ratings are routinely used to evaluate a recommender system's accuracy, by comparing how closely system-predicted ratings are able to match the users' reported ratings.

Prior studies have primarily focused on the impact of personalized item recommendations that are often presented to users as system-predicted preference ratings, e.g., predicted preference ratings for a movie on a 1-5 star scale [Adomavicius et al. 2013, 2018; Cosley et al. 2003]. Many real-world applications (such as Netflix, IMDB, and Rotten Tomatoes) display the rating-based information prominently to the user and even on the same page as the preference collection interface, This is true both where the rating is a personalized prediction or an aggregate rating from all users; but a rating is just one form of personalized item recommendation. Another common approach is to compile and display personalized lists of "best" (or recommended) items for each user, referred to as a top-N recommendation. Top-N recommender systems have been widely adopted by many e-commerce web sites to suggest lists of the best $N$ items that match customers' personal tastes, preferences, and needs (e.g., Cremonesi et al. [2010], Linden et al. [2003], Lu et al. [2015], and Smith and Linden [2017]). For example, many news websites recommend the top-10 articles to their readers based on their interests and browsing histories, Amazon.com presents personalized lists of products that their customers might find interesting, Netflix suggests lists of movies and TV shows based on their subscribers' preferences and watch histories, and Spotify generates personalized music playlists that are tailored to each individual user's music preferences. These recommendations are usually displayed on the webpage as a list of items (either ordered or not-ordered), often with no accompanying predicted rating values [Lu et al. 2015; Zolaktaf et al. 2018].

To our knowledge, no prior research has examined the impact of recommendations presented in *list-based, non-rating forms* on users' post-consumption preferences. In this study, we extend prior literature by investigating whether recommendations presented as a list of top-N items lead to bias in users' reported preference ratings. Studying this type of presentation format is highly relevant because it is among the most popular display formats used by real-world application systems. It also provides a robustness check on the biasing effects of rating-based recommendation displays that have been established in recent studies (e.g., Adomavicius et al. [2013, 2018]), as well as having practical design implications. Specifically, our study addresses the general question: *Are users' self-reported post-consumption preference ratings for items influenced by recommendations displayed in the form of a list of the top-N items?*

In addition to this general question, we recognize that top-N lists can vary as to whether the list of items is *personalized* or *non-personalized*. Personalized item lists are typically generated using recommender systems that employ collaborative filtering and/or other techniques to predict how well each unconsumed item matches with a given individual's preferences [Adamopoulos and Tuzhilin 2014; Kang et al. 2016; Loni et al. 2019; Ricci et al. 2015; Xue et al. 2019; Zolaktaf et al.

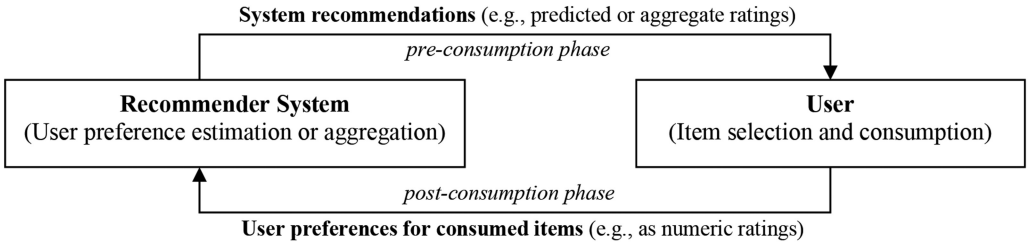


Fig. 1. Feedback loop in user-recommencder interactions (adapted from Adomavicius et al. [2013]).

2018]. Such predictions can then be used as a basis for compiling a list of items predicted to be most relevant for a given individual (e.g., "Here are the top recommendations for you" at Amazon's home page, "Top Picks for you" at Netflix, or "Discover Weekly" at Spotify); and, these personalized top-N lists are likely to be very different from user to user. On the other hand, non-personalized, general top-N lists are typically derived from aggregate judgments (i.e., some overall consumer consensus) that is the same for all users [Prawesh and Padmanabhan 2014; Zolaktaf et al. 2018]. Common examples include lists derived from aggregating sales, downloads, click data, or user opinions (e.g., "Top Box Office" at Rotten Tomatoes, "Authoritatively ranked lists of books sold in the United States" at *The New York Times*, or "Most Popular Movies" and "Top Rated Movies" at IMDb). For both personalized and non-personalized cases, the information can be presented in an identical list format; however, their underlying meanings and theoretical underpinnings are very different. Thus, we explore whether aggregate non-personalized top-N lists have different effects on users' self-reported preference ratings as compared to personalized top-N lists. In this article, we present the results of several controlled laboratory experiments, in which we manipulated the characteristics of top-N item lists presented by a recommender system to answer our central research question.

The rest of the article is organized as follows: We first provide a brief overview of related literature. Next, we describe the experiments and present our findings. The article concludes with a discussion of implications for research and practice.

## 2 BACKGROUND AND RELATED WORK

Product recommendations and recommender systems primarily focus on providing *quality*-related information about items to the user (as opposed to describing their *content* information). The recommendations are generated and presented in multiple ways (e.g., Koren and Bell [2015], Prawesh and Padmanabhan [2014], and Zhang et al. [2019]). They can be personalized, system-predicted preference ratings that indicate expectations of how much the consumer will like items (derived using methodologies from machine learning, statistics, etc.); or non-personalized recommendations, usually in the form of aggregate user ratings.

Recommender systems often operate within a feedback loop, as illustrated in Figure 1. Whether personalized or not, system recommendations are provided to users, typically to help them with their item selection decisions. In other words, recommendations represent user-system interactions in the *pre-consumption phase*, which is where the value of the recommender systems lies—recommender systems are explicitly designed to affect user behavior in item selection. The subsequent ratings that a user provides back to the system after item consumption complete an iterative "feedback" loop of the user-system interactions. The user's feedback to the system represents user-system interactions in the *post-consumption phase*. However, it is important to note that, after consumption of the product, the system's recommendations *by design* are expected not to have any impact on users' preference judgments (i.e., on users' submitted ratings for the just-consumed

items). In other words, the common assumption is that the users' submitted ratings represent an accurate and unbiased expression of users' preferences, which is further evidenced by the fact that users' submitted ratings are routinely used by the same recommender systems as additional "ground truth" to further refine the subsequent recommender system ratings. At a high level, the setup described by Figure 1 motivates our research question.

Prior research (as discussed in the following sections) clearly indicates that system recommendations (observed and appropriately used by the users pre-consumption) can significantly affect users' post-consumption preference ratings. Since this post-consumption influence is not aligned with the normative expectations of system designers (and most likely of the users either), this phenomenon has been appropriately characterized as a bias in the literature. Our general goal is to understand whether this bias, which has been demonstrated in recommender systems that use rating-based displays, also exists in the use of recommender systems with Top-N list displays.

### 2.1 Unintended Consequences of Rating-Based Recommender Systems

Much of the research in the recommender systems literature has focused on developing and improving personalization techniques for making more accurate predictions of users' preferences for items and providing more diverse, novel, and relevant recommendations (e.g., Adomavicius and Kwon [2012], Deshpande and Karypis [2004], Konstan [2004], Lu et al. [2019], Truong et al. [2007], Xue et al. [2019], and Zhang et al. [2016]). Recently, there is an increasing body of research that examines the implications of recommendations on users, items, and retailers (e.g., Belluf et al. [2012], Chen et al. [2013], Jameson et al. [2015], Jannach et al. [2015], and Wan et al. [2020]). Research has shown that recommender systems have a greater influence on users' consumption choices than peers and experts [Senecal and Nantel 2004]. While personalized recommendations help users quickly discover relevant content, such systems may also lead to unintended consequences. For example, researchers have long debated whether recommender systems create "filter bubbles" that encapsulate users in information and opinions (such as news, search results, or social network updates) that conform to and reinforce their own beliefs, views, and interests (e.g., Nguyen et al. [2014], Pariser [2011], and Resnick et al. [2013]). On the one hand, users need protection from information overload and they generally prefer to see content they feel familiar or agree with; on the other hand, there is the risk that users don't get exposure to different information or opinions that disagree with their own viewpoints, eventually isolating them in cultural or ideological bubbles [Pariser 2011]. Also, prior research has found that online recommendations can lead to a "rich-get-richer, poor-get-poorer" effect for products, resulting in a decrease in aggregate consumption diversity over time [Fleder and Hosanagar 2009; Hosanagar et al. 2014; Vargas and Castells 2014]. Additionally, recommending top-N items can lead to popularity amplification of these items and make the system susceptible to manipulations [Abdollahpouri et al. 2017; Jannach et al. 2015; Prawesh and Padmanabhan 2014].

More recently, studies have begun to explore how online recommendations can impact post-consumption preference responses and the effects of different recommendation formats upon these post-consumption biases, i.e., Figure 1's illustration of the uninintended feed-forward influence of the pre-consumption recommendations upon the post-consumption user ratings (e.g., Adomavicius et al. [2013, 2018, 2019], Cosley et al. [2003], and Jannach et al. [2015]). Beginning with personalized system-predicted ratings, prior studies have shown strong and consistent evidence that the preference ratings provided by users after item consumption are biased toward these system-generated recommendations. For example, Cosley et al. [2003] explored the effects of system-generated recommendations on user re-ratings of movies and found that users showed high test-retest consistency when no system-predicted preference rating was provided. However, when users re-rated a movie while being shown a "predicted" personalized value that was altered

upward or downward by the researchers from their system's actual prediction by a single fixed amount of one point (i.e., providing a systematically altered higher or lower prediction), users tended to give higher or lower ratings, respectively, as compared to a control group receiving the system's actual predictions. This showed that system predictions can affect users' ratings based on preference recall, for movies seen in the past and now being evaluated.

Additionally, Adomavicius et al. [2013] examined system effects in three laboratory studies using a different controlled setting. Consumer preference ratings for items were elicited at the time of item *consumption*, thereby removing possible explanations deriving from the preference uncertainty that can be present at the point of recall when trying to evaluate one's preferences for an item that may have been experienced long ago. In this setting, one's preferences should arguably be based solely on the immediate experience of the item; no uncertainty is present. Even without a delay between consumption and elicited preference, consumers' preference ratings were consistently influenced by the system-generated personalized recommendations. The effect was observed across different content domains (TV shows and jokes). And, the effect obtained whether the recommendation was seen before or after watching a TV show; so, an explanation based on priming the viewers' expectation for the upcoming experience was not supported. Consistently, the displayed system predictions, when perturbed to be higher or lower, affected the submitted consumer ratings to move in the same direction. In addition, biases were observed to be roughly linear, with the magnitude of the drift in ratings proportional to the magnitude of the perturbation of the recommendation. Furthermore, the impact of perturbations is not necessarily symmetric (i.e., when the recommendation is adjusted upward vs. downward). Specifically, a symmetric pattern of effects was observed when analyzing aggregate impact across multiple items. In contrast, different asymmetric patterns were observed when analyzing the impact of system recommendations on specific items.

Further, recent research has found that the system-generated ratings can significantly affect consumers' economic behavior with respect to the suggested items [Adomavicius et al. 2018]. Using three controlled experiments in the context of digital song purchases, the authors found strong evidence that song recommendations substantially affected participants' willingness to pay for the songs, controlling for participants' preferences and demographics. The effects persisted even when item uncertainty was reduced by forcing participants to listen to song samples prior to pricing the songs. The effect also persisted when scale compatibility issues were removed. Scale compatibility is another common explanation for biases whereby using the same scale for system-predicted ratings and user self-reported ratings creates a demand effect to increase the correspondence between stimulus and response [Tversky et al. 1988]. In the study, willingness-to-pay judgments for digital songs were expressed using a 0-99 scale, i.e., in U.S. cents, and system ratings where expressed using a typical 1-5 star scale. Thus, the effect of system recommendations is not purely an effect of reacting to a numerical value on a common scale.

### 2.2 Impact of Personalized vs. Non-Personalized Ratings

Looking beyond personalized recommendations, Adomavicius et al. [2016] compared the decision biases generated by personalized and non-personalized (i.e., aggregate) product ratings. Though they both are forms of quality information provided to users to aid the item selection process, the two forms of information are hypothesized to activate quite different mechanisms. Aggregate ratings impact judgment through social motivations grounded in the literature on social influence [Aral 2014; Krishnan et al. 2014]. Consumers are posited to engage in a form of observational learning of how to behave based on the behavior of others. In contrast, the effects of system-predicted ratings are explained in terms of information integration and processing of useful content into forming a judgment (e.g., Mussweiler and Strack [1999] and Tversky et al. [1988]). The effects are

not social in nature. Indeed, even prediction algorithms that incorporate the preferences of other users, e.g., collaborative filtering techniques [Ricci et al. 2015], often do not alert the user of such. Despite these differences, Adomavicius et al. [2016] found parallel effects in that, when shown separately, both personalized and aggregate ratings biased users' self-reported, post-consumption preference ratings for items and to a similar extent. However, consistently with their different mechanisms, when they were presented together, the parallelism ended, and personalized ratings tended to dominate in biasing users' responses. All the evidence of biases has further generated an increase of research efforts in detecting and removing biases in user ratings (e.g., Amatriain et al. [2009], Hirt et al. [2004], Joachims et al. [2017], Schnabel et al. [2016], Soll et al. [2016], Yang et al. [2018], and Zhang et al. [2017]).

Overall, the biases resulting from system recommendations on preference judgments have been shown in multiple prior studies to be robust across a variety of digital goods, settings, and conditions. However, whether personalized or non-personalized, these prior studies have mainly focused on *rating-based* information about the quality of individual items (e.g., "we predict you will rate this item as 4.3 out of 5 stars" or "the average user rating for this item is 4.3 out of 5 stars"). Presentation of the specific rating values may be driving the observed biases, possibly through anchoring effect mechanisms as proposed in the prior research (e.g., Adomavicius et al. [2013, 2016, 2018]).

In this article, our focus is still on understanding the recommendation bias, i.e., the feed-forward influence of the pre-consumption recommendations upon the post-consumption user ratings. While the previous research on recommendation bias has focused exclusively on rating-based recommendation formats, there has been no work on exploring whether the same significant biases arise with other recommendation formats. In this work, we comprehensively study the issue of recommendation biases for top-N recommendations, another important and widely used recommendation format.

### 2.3 Top-N Recommendation Formats and Positioning of Our Study

Top-N recommendation lists identify a set of N best items that will be of interest to a certain user, and the items in the list may or may not come with rating values attached [Deshpande and Karypis 2004; Loni et al. 2019; Ning and Karypis 2012; Xue et al. 2019]. Top-N recommendation displays have been in use in many real-world applications since the early years of recommender systems research (e.g., Adomavicius and Tuzhilin [2005], Deshpande and Karypis [2004], Prawesh and Padmanabhan [2014], Smith and Linden [2017], and Xia and Karypis [2011]). Recommender systems that display Top-N items often utilize computational techniques similar to those used by systems that display predicted preference ratings for a particular item, such as averaging ratings and item-based and user-based collaborative filtering (e.g., Cremonesi et al. [2010], Deshpande and Karypis [2004], and Resnick et al. [1994]), linear methods (e.g., Ning and Karypis [2012] and Zhang and Iyengar [2002]), and matrix factorization (e.g., He et al. [2016], Kabbur et al. [2013], and Kang et al. [2016]). One of the key differences between top-N systems and preference-prediction recommender systems is in the *display* of information to the user. Traditional recommendation systems often present their recommendations as scores for individual items, e.g., displayed predicted ratings on the web page of each particular item or even as part of the data collection when users submit their feedback ratings. Well-known examples include the movie recommendations based on personalized system predictions on Netflix and the product recommendations based on aggregate user ratings on Amazon. In contrast, top-N systems display multiple recommended items at a time in a list, with or without including predicted preference ratings. The top-N format has gained popularity in real-world applications. For example, Spotify presents song recommendations as a top-N list without any rating information, whereas Amazon suggests items in "Recommended for

Table 1. Demographic Characteristics of the Participants

| # Participants | 504 |
|---|---|
| % Female | 60.3% |
| Age: Mean (SD) | 21.15 (6.61) |
| % Native English Speaker | 77.8% |
| % Undergrad | 73.6% |

you" lists to their customers with accompanying aggregate rating information and MovieLens (movielens.org) suggests items in their "top picks" lists with accompanying personalized rating predictions.

The display format of recommendations represents a system design decision that may significantly impact user behavior and is, thus, our main motivation for conducting this research. If top-N lists do not reproduce the same biases that are observed with preference rating displays, then the rating displays represent at least one aspect of the individual recommendations that is creating the observed biases. Thus, the investigation of the top-N list display moves us toward identifying the underlying mechanism for the observed biases. Also, from a practical standpoint, the top-N list could be a preferable display format for recommendation settings where bias is undesired. On the other hand, if top-N lists do reproduce the biases, then the underlying mechanism arises from making a recommendation itself and is unlikely to be a feature of the ratings. Either result will help to extend our understanding of the mechanisms underlying system recommendation biases and to inform system designers about their use of recommendations.

To further our understanding, we collect data under controlled experimental conditions, investigating post-experience reactions to items recommended via Top-N lists under a variety of conditions. In all conditions, the basic methodology is consistent. In total, we conducted three studies that included thirteen different conditions. Participants were randomly assigned among thirteen groups, offering different manipulations. We begin in the next section by describing the common methodology across all groups. For ease of exposition, the discussion of the groups is then divided into three studies. For each study, we identify the research questions, hypotheses, and the treatment groups relevant to the hypotheses and the results. A general discussion section follows to conclude the article.

## 3 COMMON METHODOLOGY

### 3.1 Participants

All three studies involved the consumption and rating of jokes, so the participant population required no special characteristics. Recruited from a US college's research participant pool, 504[1] participants were paid a fixed $10 fee for completing the study. Table 1 shows the demographic features of the sample across all conditions of the between-subjects component of the design. No participants participated in more than one treatment group. The mean time for completing the study was 18.9 minutes. For the experimental task, this was ample time, and fatigue was judged as not an issue.

[1]An *a priori* power analysis for a general linear model with 12 independent variables shows that a sample size of 125 observations is required to maintain power of 0.8, with a significance level of 0.05 and detect a medium size effect for the treatments in Study 1. This value was determined using the R packages Cohen.ESC and pwr. The model is discussed below. The analyses within Studies 2 and 3 have the same or smaller number of right-hand-side variables and, thus, power analyses would yield similar sample size requirements. For the analyses in all three studies, our sample sizes are more than sufficient for the effects we are trying to identify.

### 3.2 Stimuli

The studies used jokes from the Jester Online Joke Recommender System repository, a database of jokes and preference data maintained by the University of California, Berkeley (http://eigentaste.berkeley.edu/dataset). The Jester joke database has been used in prior recommender systems research [Adomavicius et al. 2013; Goldberg et al. 2001]. Specifically, we selected a list of 100 jokes from Dataset 2 which contains 150 jokes in total. We removed jokes that were suggested for removal at the Jester website (because they were either included in the "gauge set" in the original Jester joke recommender system or because they were never displayed or rated, [Goldberg et al. 2001], jokes that more than one of the co-authors of our study identified as having overly objectionable content, and finally jokes that were the greatest in length (based on word count).

We select jokes as the stimuli in our studies and believe that jokes represent a good domain choice for several reasons. First, jokes are relevant to the participant population being sampled, primarily consisting of university students. Second, jokes are stimuli that allow each participant to respond to (i.e., to experience/consume and rate) multiple items within a manageable time period in a single laboratory session, which is an essential aspect of the research question and which adds to the robustness of the collected data. The readers' preference ratings can be gathered immediately after the reading of each joke, and there should be no uncertainty of preference due to memory effects. Third, jokes also allow us to collect users' post-consumption ratings separate from other confounding factors—e.g., economic or behavioral influences—that may come into play with products in a field setting, such as users' budgetary or time constraints. Fourth, as noted in the literature review, the biases generated by rating-based system recommendations in the form of ratings have already been shown to be robust across a variety of stimuli types (not only jokes, but also movies, TV shows, and songs) with comparable findings across all contexts. With system ratings, the form of the stimulus has not been a factor impacting the observed bias. Therefore, given that the focus of this study is to investigate the presence (or absence) of these biasing effects for a different type of recommendation (i.e., top-N recommendations), we believe the experimental findings of this study are representative and are likely to generalize beyond jokes.

### 3.3 Procedure

In all three studies, participants first read through a list of 50 randomly selected jokes from the database. These 50 jokes were randomly selected for each participant and were not the same for different participants. Participants were asked to indicate if they had seen the joke before and also rate each joke using a 5-star rating scale with half-star increments.

Next, participants saw 20 additional jokes displayed as a recommended list. The presentation of these 20 additional jokes differed across three studies. In total, Studies 1-3 include thirteen treatment groups, and participants were randomly assigned to one of these thirteen treatment groups. Table 2 summarizes the 13 between-subject conditions and the number of respondents in each condition. The 13 groups differed only in the conditions under which the 20 jokes in the second set were viewed and rated. The description of the procedures for the 20 treatment jokes is described within each study below.

After reading each joke on the recommended list, the participants were asked to indicate whether they had seen the joke prior to the experiment and also rate the joke using the same 5-point rating scale. These 20 jokes were displayed on two consecutive pages (10 jokes on each page). Since (ordered) Top N lists have an implied ranking of items, we control for the display order of these items in our data analysis.

The final task for the participants was to complete a short survey of demographic information. The same survey was used in all three studies.

Table 2. Experimental Conditions for Studies 1–3

| Group | N | Study | Recommendation Description | Actual Operationalization |
|---|---|---|---|---|
| 1 (Control: Fully Random) | 34 | 1&2 | Random | Random |
| 2 | 34 | 1 | Personalized | Random |
| 3 | 39 | 1 | Personalized | Best |
| 4 | 34 | 1 | Aggregate | Random |
| 5 | 43 | 1 | Aggregate | Best |
| 6 | 40 | 2 | Ordered Personalized | Random |
| 7 | 40 | 2 | Ordered Personalized | Best |
| 8 | 39 | 2 | Ordered Aggregate | Random |
| 9 | 42 | 2 | Ordered Aggregate | Best |
| 10 | 41 | 3 | Random | Random w/Numbers (Random#) |
| 11 | 38 | 3 | Ordered Random | Random w/Numbers (Random#) |
| 12 | 40 | 3 | Personalized | Random w/Numbers (Random#) |
| 13 | 40 | 3 | Ordered Personalized | Random w/Numbers (Random#) |

*Notes*:
1. See Appendix A for the example interfaces that the subjects saw in different treatment groups.
2. Additional descriptions of these treatment groups are provided in Sections 4-6.

## 4 STUDY 1: UNORDERED ITEM LISTS

### 4.1 Hypothesis

In identifying bias, our studies, summarized in Section 3, adopted the standard measure of comparing (i) responses when a high system-predicted rating value is provided to (ii) responses when a low rating value is provided. Consider this example typical for identifying an anchoring bias [Tversky and Kahneman 1974]: Subjects are asked to assess a quantity like the percentage of African nations in the United Nations. Before estimating this percentage, participants were required to spin a wheel labeled with numbers from 0 to 100. The wheel was rigged to land only on either 10 or 65. The exposure to this "random" number created an anchoring bias that impacted the participants' estimates. The authors report the evidence of this bias as follows: "the median estimates of the percentage of African countries in the United Nations were 25 and 45 for groups that received 10 and 65, respectively, as starting points" (p. 1128). The identification and measurement of the bias are done by showing that high and low starting points yield different estimates in a situation where the starting point should not matter; if bias does not exist, different starting points would result in the same, unbiased response on average.

For a top-N list, this exact procedure is less reasonable since it would require the display of a bottom-N list, which users (and, thus, our study participants as well) are not accustomed to receiving. This is unlike low predicted preference ratings, which are commonly observed in practice. Consequently, a different measure is employed for identifying any bias introduced from displaying a top-N list of recommended items. The essential comparison is between (a) responses made to a list of items identified as a top-N list either based on system recommendations or aggregate consumer ratings and (b) responses made to a list of items not identified as a top-N list. If users react differently when the top-N label is randomly applied or not applied to the same list of items, then any difference in subsequently submitted user preference ratings between items on the two lists is attributable to the former list being identified as a "recommended" list. Consistently with prior literature, this type of difference is interpreted as bias, because item quality information in the form of a recommendation is useful prior to consumption; but it serves

no value to the individual after the item consumption, and so is explicitly or implicitly inferred by system designers and providers to have no effect.

This measure is most comparable to a slightly different variation used in the prior studies, namely the comparison between (i) responses made when receiving a high prediction and (ii) responses made when receiving no prediction. As observed by Adomavicius et al. [2013], when measured for a single item, e.g., a single TV show, a difference between these conditions is not necessarily reliably observed. However, when measured over multiple items, Adomavicius et al. [2013] observed a marginally significant effect between High and Control conditions with TV shows. With greater sample sizes, providing greater power and using jokes as stimuli, Adomavicius et al. [2019] substantiated a statistically significant effect between High and Control conditions comparable to those we will use in the current studies.

Also, as Adomavicius et al. [2016] saw a similar bias when aggregate ratings were provided as to when system predictions were provided, despite their different psychological mechanisms, we expect parallel results for these two types of quality information in the present study. Hence, we advance the following hypothesis:

*Hypothesis 1.* Holding all else equal, the labeling of a list of items as top-N will generate higher preference responses (compared to when the list is not so labeled), both when the top-N list is presented as being generated from (a) personalized system predictions and (b) aggregate user ratings.

### 4.2 Treatment Groups

Treatment groups 1–5 (Table 2) comprise a $2 \times 2$ between-subjects factorial design with an added control group. In all five groups, the participants received 20 jokes in a list. They were explicitly told that the jokes were in no particular order. Different among the groups was what they were told about the source of the list and the procedure by which the list was generated.

In the *Fully Random* condition that serves as the control group, participants were told that the additional 20 jokes were randomly selected from an existing database and displayed in no particular order. These jokes were indeed random selections from the database, and randomly ordered for each subject, in the actual operationalization of the experiment.

For the $2 \times 2$ factorial component of the design, the first factor was whether the list was described to the participant as representing aggregate or personalized information. In the *Personalized* conditions, participants were told that the list represented the top-N jokes selected by a recommender system based on their individual preferences. In the *Aggregate* conditions, participants were told that the list represented the top-N jokes selected according to aggregate user ratings on these jokes. In both conditions, participants were explicitly told that the jokes within the list were displayed in no particular order.

The second factor in the $2 \times 2$ design was whether the 20 jokes were actually the best 20 from the database as indicated, or if the jokes were just 20 randomly generated jokes from the database. In the *Random* conditions, the 20 jokes were randomly generated. As an additional control for joke funniness, we displayed the same exact list of 20 jokes to all participants as were used in the *Fully Random* control condition, again with the order of jokes randomized for each person. In the *Best* conditions, the jokes were the actual top 20 jokes that had the highest predicted ratings for the specific participant or the highest aggregate ratings. To be consistent with the other conditions in this study and with what participants were told, the jokes within the list were presented in no particular order. The 20 jokes with the highest aggregate ratings were selected as those with the highest mean user ratings based on the Jester dataset. The 20 jokes with the highest predicted ratings were estimated using the well-known item-based collaborative filtering recommendation algorithm [Deshpande and Karypis 2004; Sarwar et al. 2001] based on the Jester dataset as

Table 3. Study 1 Treatment Groups

| Group | Treatment Group | Recommendation Description | Actual Operationalization |
|---|---|---|---|
| 1 | Fully Random | "These are 20 additional jokes from our database (in no particular order)." | A list of 20 randomly-selected jokes were displayed to all participants. *In all conditions, jokes are displayed in no particular order to the participant.* |
| 2 | Personalized-Random | "Based on the ratings you provided in the previous step, our recommender system has made predictions of your personal preferences on the remaining jokes in our database. These are 20 most recommended jokes for you (in no particular order)." | A list of 20 randomly-selected jokes were displayed to all participants. |
| 3 | Personalized-Best | | A list of top 20 jokes with the highest predicted user preference ratings were selected for each participant; each participant saw a different list of jokes. |
| 4 | Aggregate-Random | "Based on other users' ratings, we have computed average funniness for all jokes in our database. These are 20 most overall liked jokes (in no particular order)." | A list of 20 randomly-selected jokes were displayed to all participants. |
| 5 | Aggregate-Best | | A list of top 20 jokes with the highest mean user ratings were selected. |

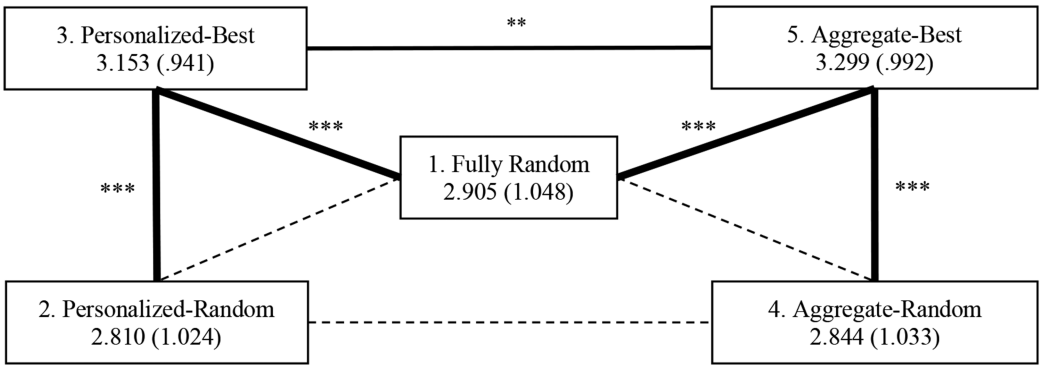


Fig. 2. Contrasts of mean user ratings by experimental conditions (1-5), Study 1.

well as the ratings data on the 50 initial jokes rated by that participant. More details about the recommendation algorithm are provided in Appendix B. The *Best* conditions were included as another form of control, allowing us to compare the effects of joke quality in a top-N list vs. any effects of bias from simply labeling the random items as a top-N list. Table 3 summarizes the recommendation descriptions and actual operationalizations for different treatment groups in Study 1. Example recommendation interfaces are provided in Figures A1–A3 in Appendix A.

### 4.3 Results

Figure 2 illustrates the pairwise *t*-tests that compare mean user submitted ratings for the experimental conditions. When random jokes were provided, identifying the jokes as recommended, either based on system predictions (*Personalized-Random*) or aggregate ratings (*Aggregate-Random*), did not lead to significantly higher user ratings compared with the *Fully Random* group. In other

Table 4. Regression Analysis, Study 1

| DV: UserRating | Coefficient (SE) |
|---|---|
| *Groups* (Fully Random as Base) | |
| Aggregate-Best | 0.075 (0.085) |
| Aggregate-Random | −0.031 (0.071) |
| Personalized-Best | 0.052 (0.091) |
| Personalized-Random | −0.071 (0.081) |
| PredictedRating | 0.902 (0.057)*** |
| DisplayPosition | 0.004 (0.003) |
| Controls | |
| ifReadBefore | 0.458 (0.271) |
| jokeFunniness | −0.013 (0.085) |
| Age | −0.002 (0.002) |
| Male | 0.044 (0.051) |
| Undergrad | 0.016 (0.074) |
| native | 0.073 (0.058) |
| Constant | 0.28 (0.225) |
| $R^2$ | 0.2405 |
| $F$ (12, 183) | 53.55*** |

*** $p < .001$; Std. Err. adjusted for 184 clusters in userID.

words, the simple labeling of the list as top-N did not introduce bias in users' submitted ratings. Including the other two treatment conditions (*Personalized-Best* and *Aggregate-Best*) in Figure 2 allows us to conclude that the consumers were not just generally insensitive. On average, participants who received actual top-N recommendations, either based on personalized predictions (*Personalized-Best*) or aggregate user ratings (*Aggregate-Best*), provided significantly higher ratings on these items than participants who received random recommendations. Note that, in this case, such rating differences are not necessarily indicative of bias in user preferences, because the jokes displayed in the *Personalized-Best* and *Aggregate-Best* conditions were likely better jokes for the participants, since they were identified using commonly used recommendation methods.

We further conduct a regression analysis to control for individual characteristics and joke heterogeneity. Given that our subjects were randomly assigned to different treatment groups and the data include repeated observations for each individual, we use a regression model with a robust estimator that clusters the standard errors by subjects. To separate item quality (i.e., inherent joke "funniness") from potential bias, we also control for the system-predicted preference ratings while testing the differences across the experimental manipulations (see Table 4). Our regression model is as follows:

$$UserRating_{ij} = b_0 + b_1(Groups_{ij}) + b_2(PredictedRating_{ij}) + b_3(Controls_{ij}) + \varepsilon_{ij}$$

The five treatment groups ($Groups_{ij}$) were entered into the model as four binary indicator variables using the *Fully Random* group as the omitted, base category. $PredictedRating_{ij}$ is the system-predicted rating for participant $i$ on joke $j$, calculated using the well-known item-based collaborative filtering algorithm [Deshpande and Karypis 2004; Sarwar et al. 2001]. The inclusion of this term provides a control for heterogeneity in expected joke preferences among participants. The collection of the ratings on the first 50 jokes in the procedure (prior to the display of any recommendations) allowed us to calculate these predicted ratings for all subjects. Appendix B provides

detailed descriptions of how these personalized preference ratings were estimated. $Controls_{ij}$ is a vector of joke- and participant-related variables. The controls included in the model were: the position of the joke in the displayed list (between 1 and 20), whether participants had read the joke before seeing it in this study (yes/no, binary), the joke's funniness (average joke rating in the Jester dataset, continuous between 1 and 5), participant age (integer), gender (binary), school level (undergrad, yes/no binary), and whether they are native speakers of English (yes/no, binary). Finally, $\varepsilon_{ij}$ represents the stochastic error of the model, which is estimated using cluster robust standard errors by participant.

After controlling for the predicted rating, the difference in user ratings between the Random group and the four other treatment groups is not significant (all $p > 0.38$). Thus, we conclude that the higher ratings submitted by *Personalized-Best* and *Aggregate-Best* groups were not due to bias, but only due to higher joke quality. In other words, recommendations labeled as top-N items did not create significant bias in users' preference ratings.

The lack of bias with top-N recommendations, either based on system predictions or aggregate ratings, contrasts with Hypothesis 1 and, more importantly, with the previously robust findings of bias generated by predicted preference rating recommendations found in the extant literature. If reliable, the result provides important implications for system design, supporting a possibly less biased recommendation mechanism. The treatments analyzed as Study 2 provide verification of the result using a stronger manipulation.

## 5 STUDY 2: ORDERED ITEM LISTS WITHOUT RATINGS

### 5.1 Hypotheses

In the recommender systems context, a top-N recommendation list can be presented as either an ordered or unordered list. We examined whether top-N recommendation lists generate bias when ordering information is added to the list, thereby increasing the list's informativeness. This element contrasts with Study 1 where the list is labeled as the best N items in no particular order. Explicitly including the order information in the top-N recommendation list is of practical interest since this is a common recommendation display design in many real-world applications (e.g., New York Times bestseller lists, Amazon.com product list recommendations, Rotten Tomatoes' list of top movies). Prior literature has consistently suggested that an online position effect exists and that rank order can significantly influence consumers' behaviors, such as click-through rates and purchase decisions, even after controlling for the quality of the displayed items (e.g., Agarwal et al. [2011], Ghose et al. [2012], Jerath et al. [2011], and Ursu [2017]). It is well recognized in both research and practice that consumers find ordered lists, such as the Billboard's music charts and U.S. News' college ranking, to be informative and influential (e.g., Isaac and Schindler [2014] and Monks and Ehrenberg [1999]). Consumers often make their decisions based on an item's inclusion in an ordered list [Sorensen 2007] or on its direction of movement on the ordered list [Pope 2009].

The basic psychological dynamic is that ordering the information in lists helps consumers differentiate the items, providing more information. In this way, the ordering of the list strengthens the manipulation by providing further emphasis on the selection of the items on the list. Further, the design allows a more sensitive test of possible bias by affording the option of testing whether participants' preference ratings vary with the ordering of the items in the list. Following the initial expectations as provided by Hypothesis 1, we hypothesize:

*Hypothesis 2.* Holding all else equal, the labeling of a list of items as an ordered list of top-N items will generate higher preference responses (compared to when the list is not labeled as a

Table 5. Study 2 Treatment Groups

| Group | Treatment Group | Recommendation Description | Actual Operationalization |
|---|---|---|---|
| 6 | Ordered Personalized-Random | "Based on the ratings you provided in the previous step, our recommender system has made predictions of your personal preferences on the remaining jokes in our database. You will be shown a list of 20 highest recommended jokes for you. These 20 jokes are ordered by their predicted preference rating beginning with the highest recommended joke first." | A list of 20 randomly-selected jokes were displayed to all participants. |
| 7 | Ordered Personalized-Best | | A list of top 20 jokes with the highest predicted user preference ratings were selected for each participant; each participant saw a different list of jokes. |
| 8 | Ordered Aggregate-Random | "Based on other users' ratings, we have computed average funniness for all jokes in our database. These are 20 most overall liked jokes. These jokes are ordered by their aggregate average rating beginning with the highest rated joke first." | A list of 20 randomly-selected jokes were displayed to all participants. |
| 9 | Ordered Aggregate-Best | | A list of top 20 jokes with the highest mean user ratings were selected. |

top-N list), both when the list is presented as being generated from (a) personalized systems predictions and (b) aggregate user ratings.

*Hypothesis 3.* The effect described by Hypothesis 2 will decrease with the order of the item in the list, i.e., the effect will be greater for items higher in the list and lesser for items lower in the list.

### 5.2 Treatment Groups

Treatment groups 6–9 (Table 2) comprise a $2 \times 2$ between-subjects factorial design. The four treatment groups within this design are similar to groups 2–5 discussed within Study 1, except that the Top-N lists were indicated as being an ordered list of top-20 jokes instead of an unordered list of jokes. For the factorial design, the first factor was whether the list presented to the participant was described as an ordered list representing aggregate or personalized information. In the *Ordered Personalized* conditions, participants were told that they are seeing an ordered list of top-N jokes selected by a recommender system based on their individual preferences. In the *Ordered Aggregate* conditions, participants were told that it was an ordered list of top-N jokes based on aggregate user ratings for these jokes. The second factor in the $2 \times 2$ design was identical to that described for Study 1, i.e., whether the shown jokes were, in fact, a set of *Random* jokes or a set of actual *Best* jokes that had the highest predicted ratings for the specific participant or the highest aggregate ratings. Table 5 summarizes the recommendation descriptions and actual operationalization for different treatment groups in Study 2. Example recommendation interfaces are provided in Figures A4–A5 in Appendix A.

### 5.3 Results

Figure 3 illustrates the pairwise *t*-tests that compare mean user-submitted ratings for the four experimental conditions with Ordered recommendations and the *Fully Random* condition (Group 1). When random jokes were provided, identifying the jokes as recommended, either based on system predictions (*Ordered Personalized-Random*) or aggregate ratings (*Ordered Aggregate-Random*), did not lead to significantly higher user ratings compared with the *Fully Random* group. In other words, the top-N presentation format with explicit order information did not introduce bias in users' submitted ratings (not supporting Hypothesis 2).

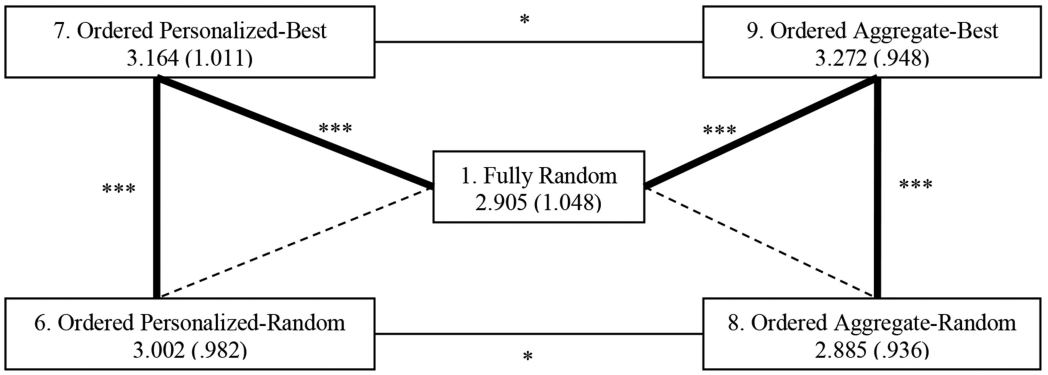


Fig. 3. Contrasts of mean user ratings by experimental conditions (1, 6-9), Study 2.

Similar to Study 1, we observed that, on average, participants who received actual top-N lists, either based on personalized (*Ordered Personalized-Best*) or aggregate ratings (*Ordered Aggregate-Best*), provided significantly higher ratings on these items than those in participant groups who received random recommendations. We conducted a similar regression analysis as in Study 1, controlling for joke funniness and predicted joke ratings, to separate item quality from potential bias and clustered the standard errors by participant:

$$UserRating_{ij} = b_0 + b_1(Groups_{ij}) + b_2(PredictedRating_{ij}) + b_3(DisplayPosition_{ij}) + b_4(Controls_{ij}) + \varepsilon_{ij}$$

The treatment groups ($Groups_{ij}$) were entered into the model as four binary indicator variables using the *Fully Random* group as the omitted, base category. Similar to the model used in Study 1, $PredictedRating_{ij}$ is the system-predicted rating for participant $i$ on joke $j$, and $Controls_{ij}$ is a vector of joke and participant-related variables. $DisplayPosition_{ij}$ indicates the position of the joke in the ordered list, ranging from 1 (highest recommended joke) to 20 (least recommended of the 20 in the top-N list). The significance of this factor would serve as evidence for Hypothesis 3. Table 6 presents the results.

After controlling for predicted rating and joke funniness, the difference in user ratings between the *Fully Random* group and the four other treatment groups is not significant (all $p > .35$). The primary driver of user's submitted ratings for jokes presented in top-N format was the estimated user preference ratings on the jokes ($b = 0.87$, $p < 0.001$). Thus, as with Hypothesis 1 and consistent with Study 1, Hypothesis 2 is not supported. There is no indication of bias even when the top recommendations are emphasized by providing an indication that the list is ordered.

Hypothesis 3 is also not supported: The displayed position of items did not show a significant effect ($b = 0.002$, $p = 0.526$). To emphasize the point, Figure 4 plots the average user rating on items by the items' displayed position. The plot confirms that the users' submitted ratings cannot be established as being affected by the items' displayed positions. Additional analyses focusing on the impact of only the very top recommendations within the lists, i.e., top-2 and top-5 items, show similar results, i.e., neither top-2 nor top-5 recommendations exhibit biasing effects, and similarly there are no item position effects within the top-2 and top-5 lists. These additional regression results are provided in Appendix C.

Combining the results from the two studies, we conclude that the higher ratings submitted by *Personalized-Best* and *Aggregate-Best* groups, whether *Ordered* (Figure 3) or not (Figure 2), were not

Table 6. Regression Analysis, Study 2

| DV: UserRating | Coefficient (SE) |
|---|---|
| Groups (Fully Random as base) | |
| Ordered Aggregate-Best | 0.112 (0.084) |
| Ordered Aggregate-Random | −0.018 (0.076) |
| Ordered Personalized-Best | −0.099 (0.085) |
| Ordered Personalized-Random | 0.054 (0.077) |
| PredictedRating | 0.87 (0.056)*** |
| DisplayPosition | 0.002 (0.002) |
| Controls | |
| ifReadBefore | 0.077 (0.048) |
| jokeFunniness | 0.032 (0.091) |
| Age | −0.006 (0.005) |
| Male | 0.106 (0.048)* |
| Undergrad | −0.038 (0.072) |
| native | −0.008 (0.058) |
| Constant | 0.441 (0.268) |
| $R^2$ | 0.2309 |
| $F(12, 194)$ | 48.61*** |

***p < .001, * p < .05, Std. Err. adjusted for 195 clusters in userID.

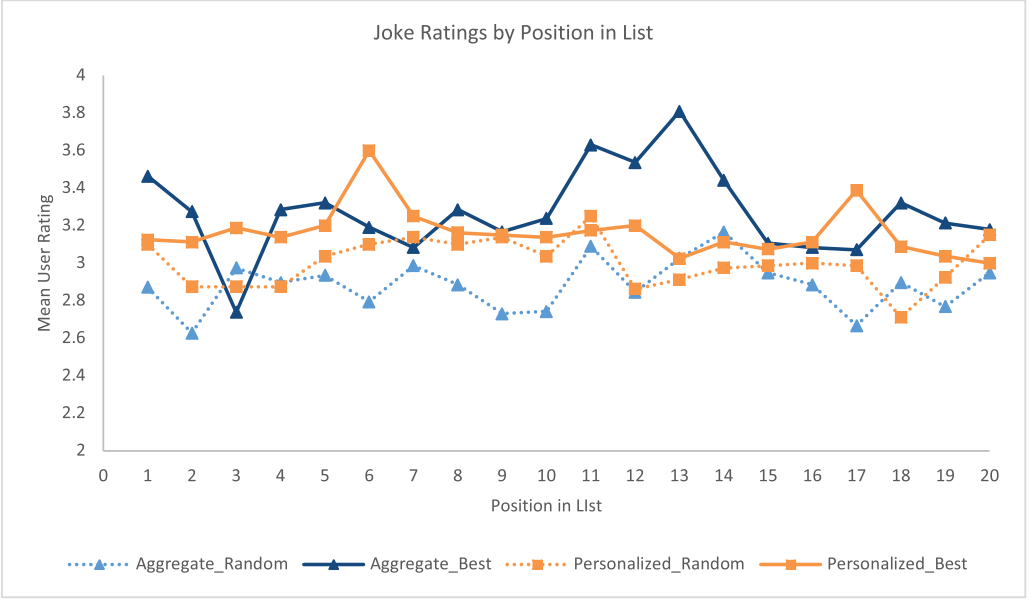


Fig. 4. Plots of mean user ratings by display position for Study 2 (all conditions have Ordered presentations).

due to bias but due to better joke quality. For randomly labeled jokes, recommendations displayed as top-N items with or without order information did not create significant bias in users' preference ratings. And, within an ordered list, there is no evidence of an effect on preference ratings for any of the recommended items whether identified at the top of the recommended list or later

in the list. These results provide further evidence that the lack of biasing effects for top-N list recommendations is robust.

As a final test, we investigate whether some aspect of the design, and its differences from prior research, is preventing us from identifying an operative bias with top-N list recommendations. Study 3 describes conditions for which the system-predicted preference ratings (i.e., numerical rating values) are added along each recommended item in the list as a check on the continuance of the results in the prior literature within the current design.

## 6 STUDY 3: ORDERED ITEM LISTS WITH RATINGS

### 6.1 Hypotheses

The results described for Studies 1 and 2 both suggested that presenting recommended items as a top-N list does not introduce bias to user's submitted preference ratings for those items. However, such biases were consistently reported in prior literature when recommendations were displayed along with predicted preference ratings—see, for example, the various predicted rating displays investigated in [Adomavicius et al. 2013, 2019]. The design within Study 3 combines the top-N list format with the numeric rating display to test whether the presentation of a top-N list *with* the associated numerical values generates bias in users' ratings. If the bias is again observed when numerical ratings are provided, consistent with prior literature, then this supports the indication that top-N lists (without the numerical ratings) are a means of providing recommendations that do not induce bias in user preference judgments. Further, it provides evidence that the display of the specific rating is the main mechanism driving the observed bias. If the bias is not observed, however, this provides evidence that the null results within Studies 1 and 2 were due to power or experimental design issues. Since Studies 1 and 2 showed similar results across aggregate ratings and personalized recommendations, we focus only on personalized recommendations in Study 3 for parsimony of design.

Prior research has shown that user preferences are biased toward the numerical rating provided in a recommendation; i.e., predicted ratings that are low will bias the users' reported preference rating to be lower than if the predicted ratings are high [Adomavicius et al. 2013, 2018]. Given these earlier robust results across a variety of settings, we propose that including numerical recommendations along with the listed items will result in biased responses. The biasing effect is tested for both unordered and ordered lists to provide parallels to the designs of Studies 1 and 2, respectively, thereby affording a comprehensive robustness check of the results for both studies.

To maintain the realism of our experiment and to reinforce the meaning of the top-N list for our participants, the numeric system ratings provided with our top-N List will be considered "high" and taken from the upper end of the numerical scale. Consequently and in contrast to the earlier research, the bias is not tested by comparing responses to high and low predicted recommendations. Instead, we expect the bias generated from the presentation of these ratings to move the participants' reported preferences higher as compared (a) to a random list of additional jokes not identified as a top-N list or, based on the failure to observe a bias in Studies 1 and 2, (b) to a top-N list without numeric ratings attached to the items in the list. The two comparisons form the hypotheses of interest:

*Hypothesis 4.* Including high numerical system ratings in a list of items will generate higher preference responses from the participants (compared to a random list of items), regardless of whether the list is unordered or ordered.

*Hypothesis 5.* Including high numerical system ratings in a list of items will generate higher preference responses from the participants (compared to a top-N list of items), regardless of whether the list is unordered or ordered.

Table 7. Study 3 Treatment Groups

| Group | Treatment Group | System Description | Actual Operationalization |
|---|---|---|---|
| 10 | Random-Random# | You will be shown a list of additional 20 jokes from our database. These 20 jokes are displayed in no particular order. | Random jokes with random ratings: A list of 20 randomly-selected jokes were displayed to all participants. All predicted rating values are randomly generated and ranged between 3 and 5. All subjects saw the same jokes and rating values. The rating values are ordered from highest to lowest for groups 11 and 13. |
| 11 | Ordered Random-Random# | You will be shown a list of additional 20 jokes from our database. These 20 jokes are ordered by their predicted preference rating beginning with the highest recommended joke first. | |
| 12 | Personalized-Random# | Based on the ratings you provided in the previous step, our recommender system has made predictions of your personal preferences on the remaining jokes in our database. These are 20 most recommended jokes for you. | |
| 13 | Ordered Personalized-Random# | Based on the ratings you provided in the previous step, our recommender system has made predictions of your personal preferences on the remaining jokes in our database. You will be shown a list of 20 highest recommended jokes for you. These 20 jokes are ordered by their predicted preference rating beginning with the highest recommended joke first. | |

### 6.2 Treatment Groups

Treatment groups 10–13 (Table 2) comprise a $2 \times 2$ between-subjects factorial design. The four treatment groups within this design are similar to groups 1, 2, and 6 discussed within Studies 1 and 2. Unlike the previous designs described as Studies 1 and 2, the manipulation was purely in how the list was presented to the participants. For all four treatment conditions, participants were shown the same list of randomly selected jokes along with the same set of numeric ratings, described as a *Random#* manipulation to contrast with the *Random* manipulation for groups 1, 2, 4, 6, and 8. These numeric ratings were randomly-generated numbers that range between 3 and 5 stars (on a five-star rating scale) to represent "high" ratings for the user.[2] In other words, all participants received the same set of information (both jokes and artificial ratings).

The variation in treatments across groups 10–13 is in the labeling of the presented list. For the factorial design, the first factor was whether the participants were told that these were 20 additional randomly generated jokes (*Random* conditions); or, in the *Personalized* conditions, they were told that these were a list of top-N jokes selected by a recommender system based on their individual preferences. The second factor in the design was whether or not the list was described as an ordered list (*Ordered* conditions). Table 7 summarizes the recommendation descriptions and actual operationalization for different treatment groups in Study 3. Example recommendation interfaces are provided in Figures A6-A9 in Appendix A.

### 6.3 Results

Figure 5 illustrates the pairwise *t*-tests that compare mean user submitted ratings for the four experimental conditions versus the *Fully Random* condition (Group 1). All four experimental

[2]Common 5-star rating scales used in practice have the midpoint of 3 correspond to a still-positive rating, e.g., "Like it".

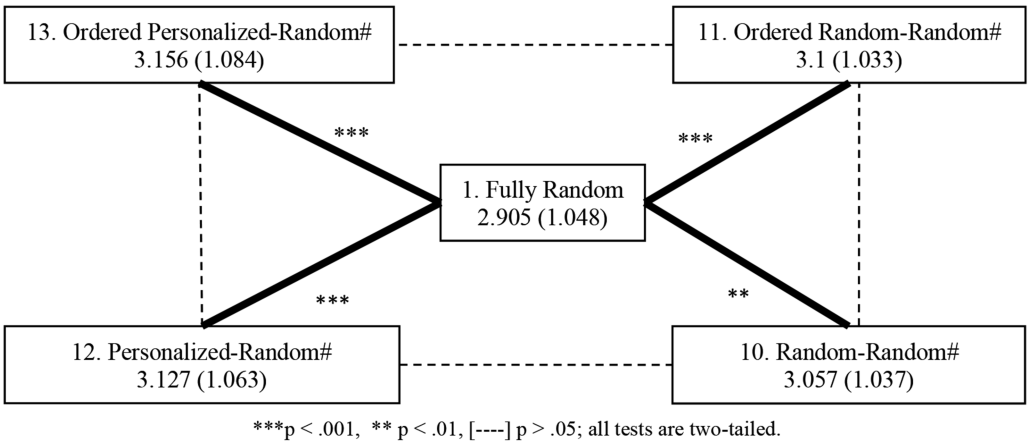


Fig. 5. Contrasts of mean user ratings by experimental conditions (1, 10-13), Study 3.

conditions have higher mean user ratings than the *Fully Random* condition. Hence, displaying the "high" numerical rating information still introduced bias, as observed in prior literature and providing support for Hypothesis 4.

In addition, none of the comparisons among the four *Random#* conditions was significant, indicating that users' ratings are comparable when numerical recommendations are provided, even though they were given different information about the nature of the list. The lack of a bias effect due to labeling the list as top-N Personalized list or due to ordering the list confirms the results described for Studies 1 and 2. More specifically, consistent with Study 1 and in contrast to Hypotheses 1a and 2a, presenting items as top-N personalized recommendations did not introduce any bias in users' submitted ratings when compared with presenting items as random items (i.e., *Personalized* vs. *Random* conditions) whether the list was in unordered form (H1a) or in ordered form to accentuate the item differences (H2a). Together, these results provide evidence that the observed null effects in Studies 1 and 2 are robust. By observing significant bias from the presence of numerical ratings, but not from the labeling or ordering of the lists, we have further support that top-N lists without numerical ratings can be an effective mode for presenting recommendations without inducing biases.

As a final analysis and in order to test Hypothesis 5, we construct the design using the non-*Aggregate* conditions for which participants received *Random* jokes, either with or without the numerical values, i.e., groups 1–2, 6, 10–13 (Table 2). These seven conditions form a three-factor 2 × 2 × 2 between-subjects design with one missing cell, as illustrated in Table 8. The three binary factors of the design are:

1) *Ordered*: The items in the list are described (and presented) as either in order of decreasing predicted rating (1) or in random order (0).
2) *Personalized*: The list is described as either a list of the top 20 personalized recommended items (1) or a random list of additional jokes (0).
3) *Rating#*: The numerical ratings are either shown for each item (*Random#*, coded as 1) or not (*Random*, coded as 0).

The missing cell in the experimental design arises from the *Fully Random* condition in which the list is portrayed as random and with no numerical ratings provided. Ordering the list by the magnitude of the personalized ratings (the missing cell) does not make sense in this situation, and

Table 8. Between-subjects Experimental design using *Personalized* data conditions from Table 2

| With Rating Numbers / Without Rating Numbers | Items Ordered | Items Unordered |
|---|---|---|
| **List Portrayed as Random** | 11. Ordered Random-Random# / [empty] | 10. Random-Random# / 1. Fully Random (Random-Random) |
| **List Portrayed as Personalized Top-N** | 13. Ordered Personalized-Random# / 6. Ordered Personalized-Random | 12. Personalized-Random# / 2. Personalized-Random |

so that cell is appropriately left missing. Since we are hypothesizing about (and are interested in) main effects and certainly have no basis for positing a three-way interaction among the factors, the missing cell does not hinder the analysis.

Pooling the data in this way allows us to test each of the main effects for these factors. The factor *Rating#* in the design offers a confirmation of the robust result from prior literature within the current setting. A significant effect of adding the numerical information provides support for the hypothesis that the bias identified in Figure 5 arises from the numerical ratings (Hypothesis 5), while controlling for various other variables.

With respect to the main effect of ordering the top-N list, the comparison of interest is similar: Does ordering the list, compared to not ordering the list, induce a bias in user responses? Note that the question is distinct from that of Hypothesis 2a which tested whether what was presumed to be a stronger manipulation of a ordered top-N list would lead to a bias compared to a random list. In this analysis, we are testing the effect of ordering itself as a factor (i.e., the *Ordered* factor), comparing between an ordered and an unordered list. Given the lack of support for Hypothesis 2a, however, we do not expect a main effect of the Ordered factor in the current analysis. Finally, with respect to the main effect of identifying the list as a top-N *Personalized* list, testing for this main effect parallels the test done in Study 1 of Hypothesis 1a.

We ran the following regression model using robust standard errors clustered by participant and controlling for the predicted ratings and item rank order (Model 1 in Table 9):[3]

$$UserRating_{ij} = b_0 + b_1(Rating\#_i) + b_2(Ordered_i) + b_3(Personalized_i) + b_4(PredictedRating_{ij}) + b_5(DisplayPosition_{ij}) + b_6(DisplayPosition_{ij} * Ordered_i) + b_7(Controls_{ij}) + \varepsilon_{ij}$$

Similar to the prior regression analyses, $PredictedRating_{ij}$ is the system-predicted rating for participant *i* on joke *j*, $DisplayPosition_{ij}$ indicates the position of the joke in the ordered list for participant *i* on joke *j*, and $Controls_{ij}$ is a vector of joke and participant-related variables. Since the order position is only explicitly identified in the *Ordered* conditions, we also add the interaction term $DisplayPosition_{ij}$*$Ordered_i$ in case the display order were to only have an effect when highlighted, i.e., in the *Ordered* conditions.

The column for Model 1 in Table 9 presents the results of our regression analyses. Of initial interest is the statistically significant relationship of the user-submitted post-consumption ratings with the *Rating#* factor in Model 1. In other words, when numerical predicted ratings accompany

[3]For the models in Table 9, we also ran the models including the two-way interaction terms for the factorial design. All interaction terms were not statistically significant (all $p > .15$) and are not included in the table.

Table 9. Regression Analysis of Pooled Data, Studies 1-3 (Robust Standard Errors in Parentheses)

| DV: UserRating | Model 1 (Full Factorial) | Model 2 (Without Ratings) | Model 3 (With Ratings) | Model 4 (With Ratings) |
|---|---|---|---|---|
| *Factors* | | | | |
| Rating# (=1) | 0.194 (0.043)*** | | | |
| Ordered (=1) | 0.178 (0.063)** | 0.050 (0.098) | 0.237 (0.079)** | 0.016 (0.091) |
| Personalized (=1) | −0.011 (0.046) | −0.085 (0.079) | −0.001 (0.057) | −0.001 (0.057) |
| PredictedRating | 0.815 (0.051)*** | 0.758 (0.070)*** | 0.864 (0.071)*** | 0.864 (0.072)*** |
| DisplayPosition | 0.002 (0.003) | −0.002 (0.004) | 0.006 (0.003) | 0.005 (0.003) |
| DisplayPosition × Ordered | −0.016 (0.004)*** | 0.0001 (0.006) | −0.026 (0.006)*** | −0.005 (0.006) |
| ShownRating | | | | 0.211 (0.040)*** |
| *Controls* | | | | |
| ifReadBefore | 0.262 (0.047)*** | 0.047 (0.082) | 0.340 (0.053)*** | 0.335 (0.053)*** |
| jokeFunniness | 0.168 (0.071)* | 0.151 (0.100) | 0.167 (0.097) | 0.165 (0.096) |
| Age | −0.001 (0.003) | −0.0003 (0.055) | −0.0001 (0.004) | −0.0001 (0.004) |
| Male | 0.039 (0.043) | 0.055 (0.062) | 0.039 (0.061) | 0.039 (0.061) |
| Undergrad | 0.113 (0.060) | 0.111 (0.088) | 0.153 (0.080) | 0.153 (0.080) |
| Native | −0.069 (0.059) | −0.063 (0.077) | −0.081 (0.087) | −0.081 (0.087) |
| Constant | −0.037 (0.169) | 0.218 (0.245) | −0.049 (0.217) | −0.881 (0.256)*** |
| $R^2$ | 0.269 | 0.227 | 0.293 | 0.299 |
| $F$ | $F(12, 266) = 99.76$, $p < .0001$ | $F(11, 107) = 35.61$, $p < .0001$ | $F(11, 158) = 74.89$, $p < .0001$ | $F(12, 158) = 71.12$, $p < .0001$ |

***p < .001, ** p < .01, * p < .05; Std. Err. adjusted for 267 (Model 1), 108 (Model 2) and 159 (Models 3 and 4) clusters in userID.

items in a Top-N personalized list, compared to when they do not, higher preference responses are obtained. Thus, the results are consistent with the robust findings of prior literature that users are biased by numeric ratings when they are asked to rate individual items. The overall design is sufficiently sensitive in terms of statistical power to detect the usual bias observed in reaction to system recommendations, even for the different measure used for bias with top-N lists.

The lack of a statistical effect for the *Personalized* factor provides additional evidence consistent with Study 1. There is no evidence that simply identifying a list as a top-N list (vs. just as another set of items) impacts post-consumption user ratings.

In contrast to Study 2, however, looking across several factors seems to indicate that ordering the list can have an effect. For Model 1, there is a statistically significant *DisplayPosition* × *Ordered* interaction effect. The main effect of the *Ordered* factor is also statistically significant; but, this must be interpreted in light of the interaction. This appears to contradict the result in Study 2 where no order effect was found. However, Study 2 only involved conditions where numerical system-predicted ratings were not included with the items. So, to understand the relationship further, we repeat the analysis described by Model 1 separately for conditions where the numerical ratings did not accompany the items and conditions where the ratings accompanied the items. Specifically, Models 2 and 3 (Table 9) were fit identically to Model 1, except for removing Rating# as a predictor. Instead, we segregate the analyses by this predictor: Model 2 is fit using only the conditions where the numerical system-predicted ratings did not accompany the items (i.e., where Rating# = 0); Model 3 is fit using only the conditions where the numerical ratings did accompany the items (Rating# = 1).

As is clearly observed, the effect of order on user's responses only occurred when the system prediction in the form of numerical ratings accompanied the items in the top-N list (Model 3). Without these numerical system ratings, no significant main or interaction effect was observed for the order of items. To bring the point home further, we re-run Model 3, but now including the system-predicted ratings as a variable in the analysis. The fitted model is:

$$\begin{aligned} UserRating_{ij} = b_0 + b_1(Ordered_i) + b_2(Personalized_i) + b_3(PredictedRating_{ij}) \\ + b_4(DisplayPosition_{ij}) + b_5(DisplayPosition_{ij} * Ordered_i) + b_6(ShownRating_{ij}) \\ + b_7(Controls_{ij}) + \varepsilon_{ij} \end{aligned}$$

where $ShownRating_{ij}$ is the displayed rating for participant $i$ on joke $j$. Again, standard errors are clustered by participant.

As observed for Model 4 in Table 9, adding the shown system-predicted rating as a variable in the analysis results in the disappearance of the ordering effect, while the predicted rating (i.e., estimated user's preference for a joke) and the actual shown numerical rating (i.e., the system's recommendation) are related to the users' responses. Hence, there is no systematic order effect, consistent with the results of Study 2. The only statistically significant effect observed in the data is the effect of numerical ratings (as part of the system's recommendation) on users' post-consumption responses, consistent with this robust finding in past studies.

## 7 DISCUSSION

### 7.1 Theoretical Contribution

The lack of significant bias in user preference ratings from top-N lists presents an interesting contrast to prior work in this area. Specifically, prior research has shown robustly that the system recommendations presented as item-specific predicted ratings consistently impact users' self-reported preference ratings, which can have deleterious effects [Adomavicius et al. 2013; Cosley et al. 2003]. We have tested a common alternative information display for item recommendations that use top-N lists to signal item-quality-related information. Importantly, the use of top-N lists to present item recommendations seems to greatly reduce, if not completely remove, the biases observed in prior studies. However, this reduction is only observed when the list items are not accompanied by the numerical system recommendations.

The results shed light on the possible mechanisms underlying the generation of decision biases from item recommendations. The lack of bias from labeling recommended item lists as top-N suggests that simply identifying items as a good match for a user is not enough to induce bias. We find that the presentation of numerical values of system-predicted preferences generated biases similar to those observed in prior research. The presence of bias with the display of numeric ratings along with the failure to observe bias when items are simply identified as recommended strongly suggests that the user preference bias is generated from some aspect of the individual item information that is presented. An important comparison to this finding are the results of Adomavicius et al. [2019]. Those researchers found that biases could be reduced, but not eliminated, by using graphical displays or thumbs up/down displays, i.e., non-numerical system-predicted ratings as recommendations. In other words, removing the number from the display helped to reduce the bias. In the present studies, we observe that removing the individual rating information completely and presenting the recommendations as a list reduces any bias to a non-significant level at most. This suggests that the biasing effect arises primarily or exclusively from individuating information about the items being recommended.

Also of interest are a couple of secondary findings. Studies 1 and 2 found evidence corroborating earlier findings of Adomavicius et al. [2016]. They found decision biases of similar magnitude in response to recommendations generated by personalized system predictions vs. non-personalized,

aggregate product ratings. These two forms of quality information are theoretically tied to different mechanisms: information integration processes and social influences, respectively. As in their study, however, we observed comparable effects of these two forms of recommendation when presented separately.

Finally, we observe no significant ordering effect either at the top of the list (top-2 or top-5) or throughout the full list up to the common value used throughout all condition of N = 20. Any ordering effect that is observed only arises when numerical recommendations accompany the items, and is fully attributable to the differences in the numeric ratings, and not the ordering itself.

### 7.2 Practical Implications and Future Work

There are a number of obvious implications for the design of recommender system interfaces and online retail displays. The primary goal of recommender systems is to reduce the cognitive load and search costs of consumers as they identify products and items to consider. When recommender systems generate biases in user preferences, it is a concern for system designers, retailers, and consumers, since a trade-off between preference bias and recommendation usefulness is created. The results of the studies presented here suggest that unordered top-N lists without numeric predictive ratings provide a means of delivering that value to consumers without the potential downside of inducing decision biases. Online retailers and platforms concerned with preference bias from their recommendations should consider the use of top-N lists.

Additionally, minimizing the display of numeric information along with recommendations should go a long way to reduce preference biases. Through usability and A/B testing, system developers can adjust several aspects of the top-N list design to maximize its usefulness for users, including but not limited to the list size, ordering, information content, and layout. Furthermore, the neutral effect of the top-N format on user preferences may make it preferable for use in the display of personalized content and information in other contexts, a promising area for a more targeted investigation.

Our findings open up a couple of directions for future research. First, while "student subjects do *not* intrinsically pose a problem for a study's external validity" [Druckman and Kam 2011, p. 41], replicating the current studies with broader populations would provide further evidence for generalizability. Second, it would be beneficial to extend this line of work beyond the joke stimuli used in this study. Showing that certain top-N recommendation formats do not exhibit biasing effects even in one application domain (jokes) is a novel and important finding, as biasing effects of rating-based recommendation formats have been consistently and robustly demonstrated for a wide variety of domains, including jokes, TV shows, movies, and songs. However, a deeper exploration of top-N recommendation effects for different product types–for instance, experiential/taste-based products vs. utilitarian/functional products–represents an important direction for future work.

## 8 CONCLUSION

Using controlled laboratory experiments, we show that top-N-list recommendation formats are advantageous for displaying item recommendations in that they do not induce bias in users' post-consumption preference judgments. This result is found to be robust for both lists of personalized recommendations and lists of top-rated items on average. Adding numerical predicted ratings to the listed items causes the robust biases seen in earlier work (i.e., for rating-based recommendation formats) to re-arise. In contexts where preference biases are of concern to an online retailer or platform, top-N lists, without numerical predicted ratings, appear to be a good choice for displaying item recommendations.

# APPENDICES

## A USER INTERFACE FOR DIFFERENT TREATMENT GROUPS USED IN THE STUDIES

Appendix A provides screenshots of user interfaces for different treatment groups across three studies. Table A1 presents a summary of all these figures.

Table A1. Summary of Figures and Treatment Group Descriptions

| | |
|---|---|
| Figure A1. | Study 1 & 2: Fully Random, unordered, no predicted rating |
| Figure A2. | Study 1: Personalized, unordered, no predicted rating |
| Figure A3. | Study 1: Aggregate, unordered, no predicted rating |
| Figure A4. | Study 2: Personalized, ordered, no predicted rating |
| Figure A5. | Study 2: Aggregate, ordered, no predicted rating |
| Figure A6. | Study 3: Random with predicted ratings, unordered |
| Figure A7. | Study 3: Random with predicted ratings, ordered |
| Figure A8. | Study 3: Personalized with predicted rating, unordered |
| Figure A9. | Study 3: Personalized with predicted rating, ordered |

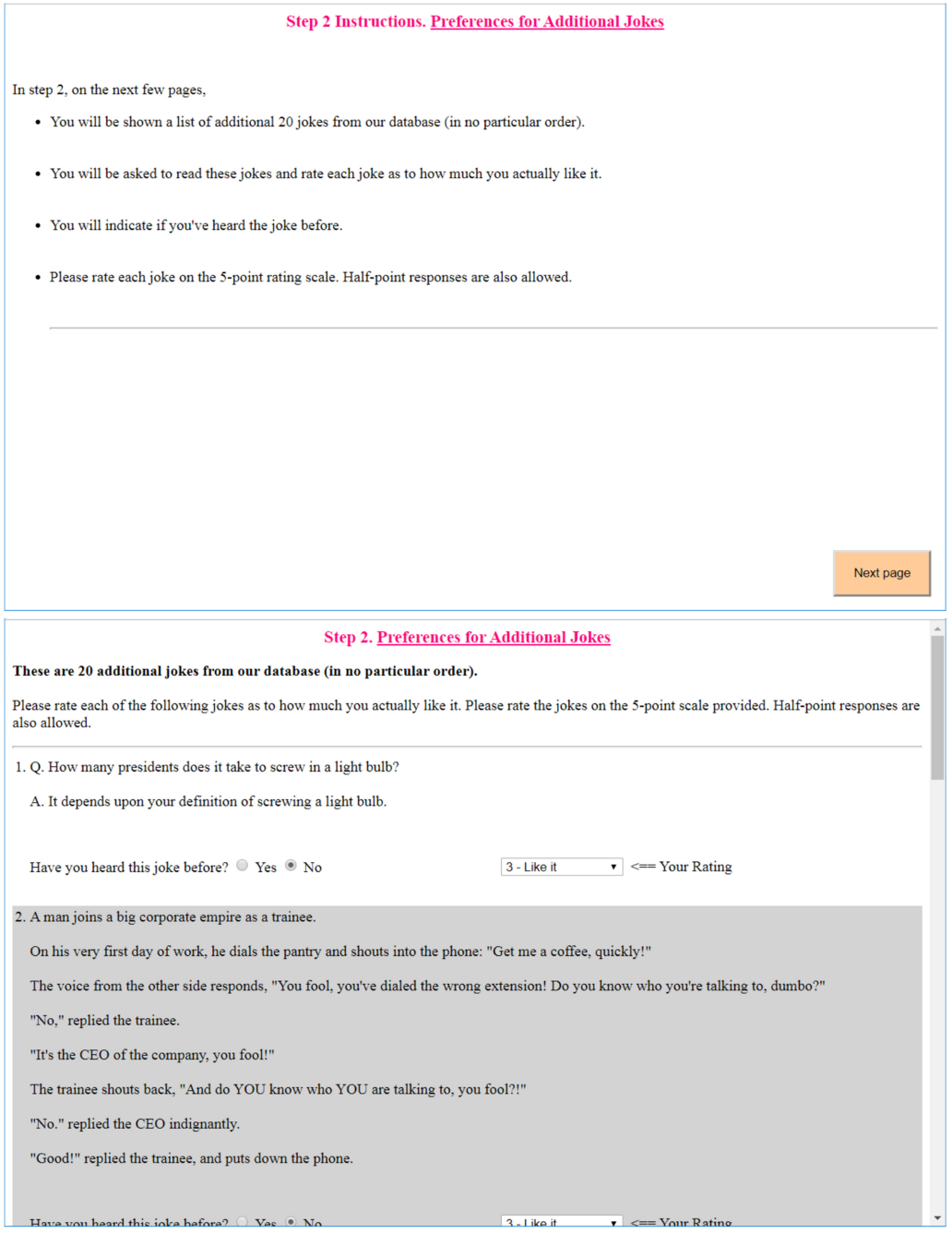


Fig. A1. Study 1 & 2: Fully Random, unordered, no predicted rating.

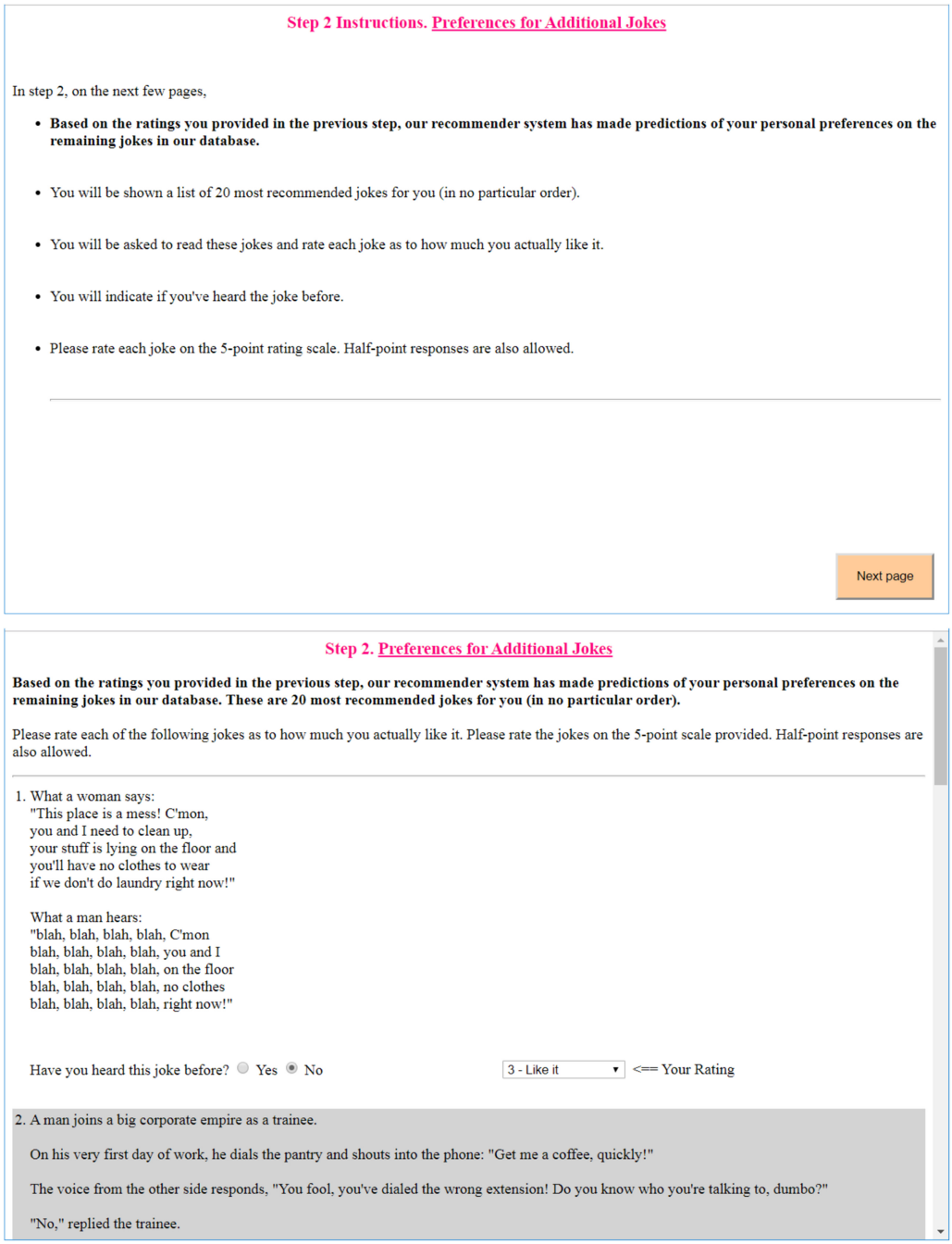


Fig. A2. Study 1: Personalized, unordered, no predicted rating.

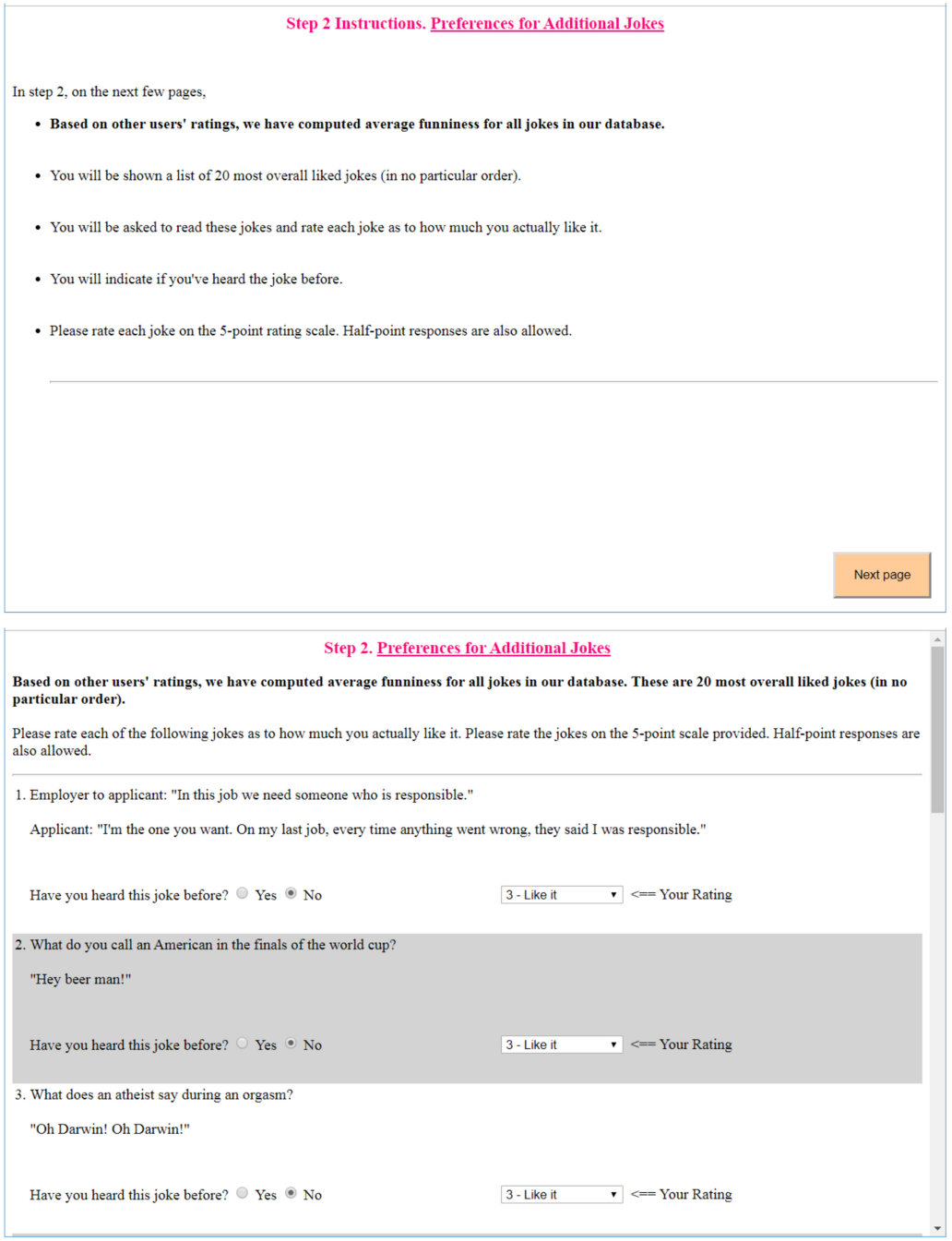


Fig. A3. Study 1: Aggregate, unordered, no predicted rating.

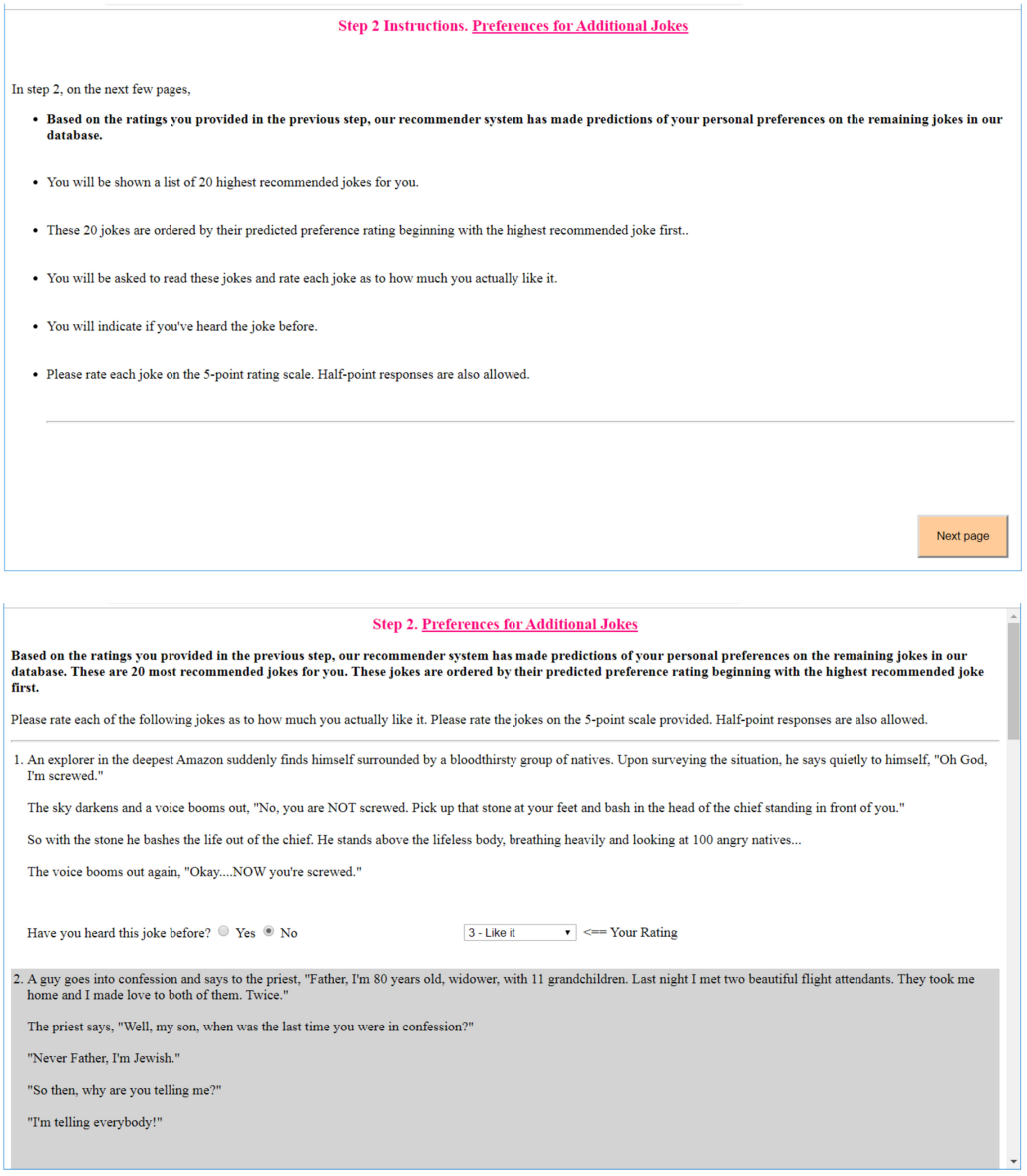


Fig. A4. Study 2: Personalized, ordered, no predicted rating.

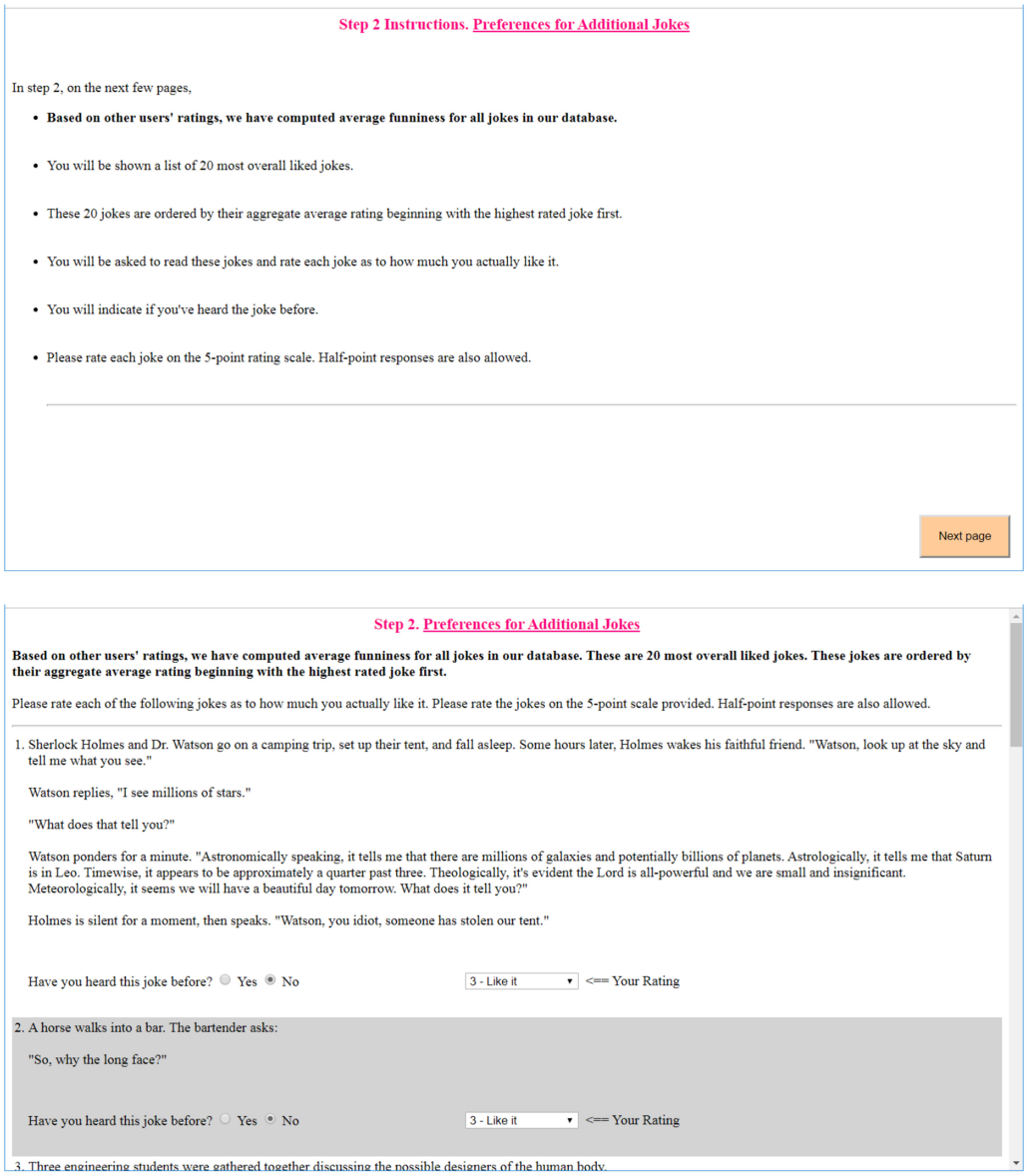


Fig. A5. Study 2: Aggregate, ordered, no predicted rating.

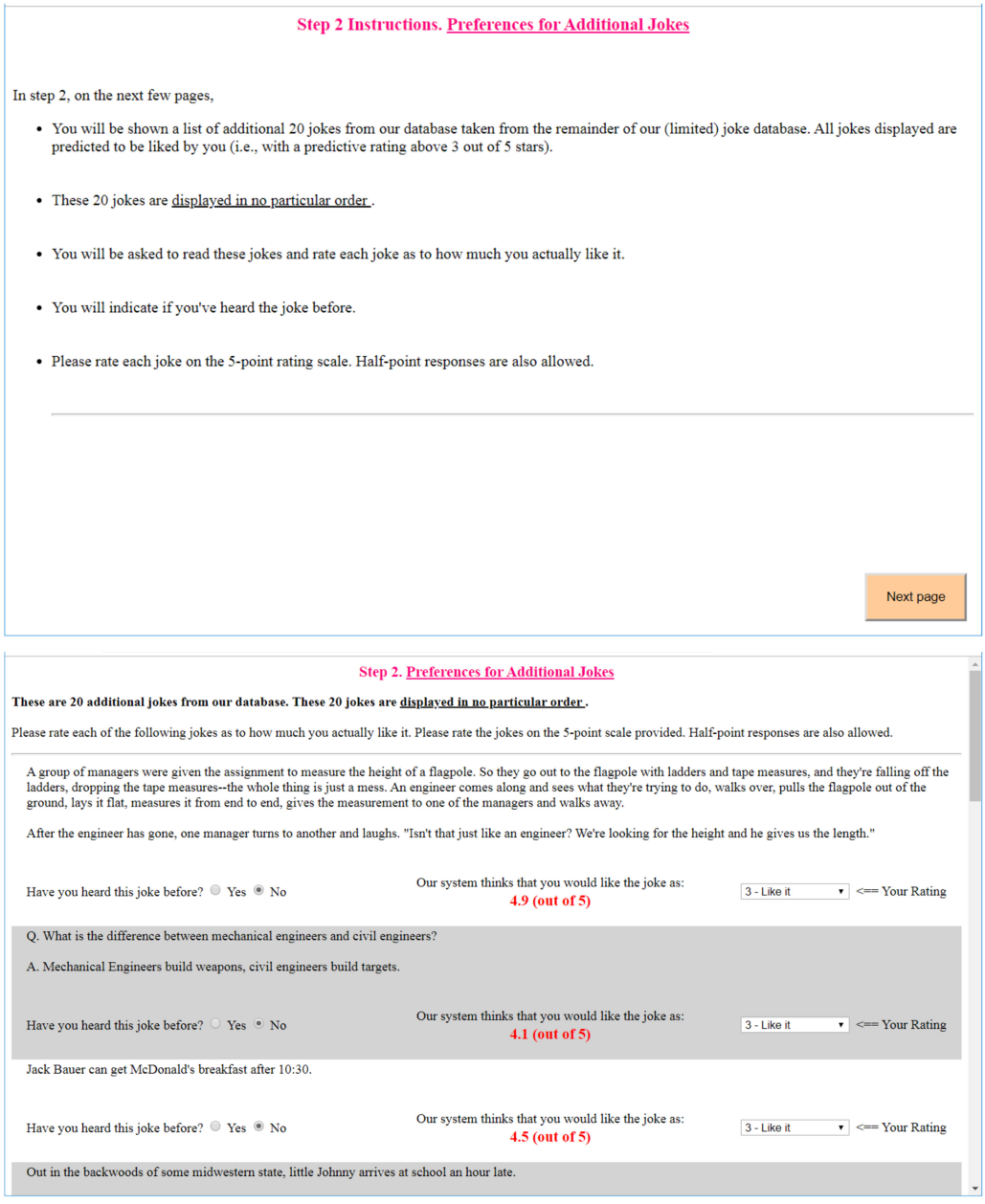


Fig. A6. Study 3: Random with predicted ratings, unordered.

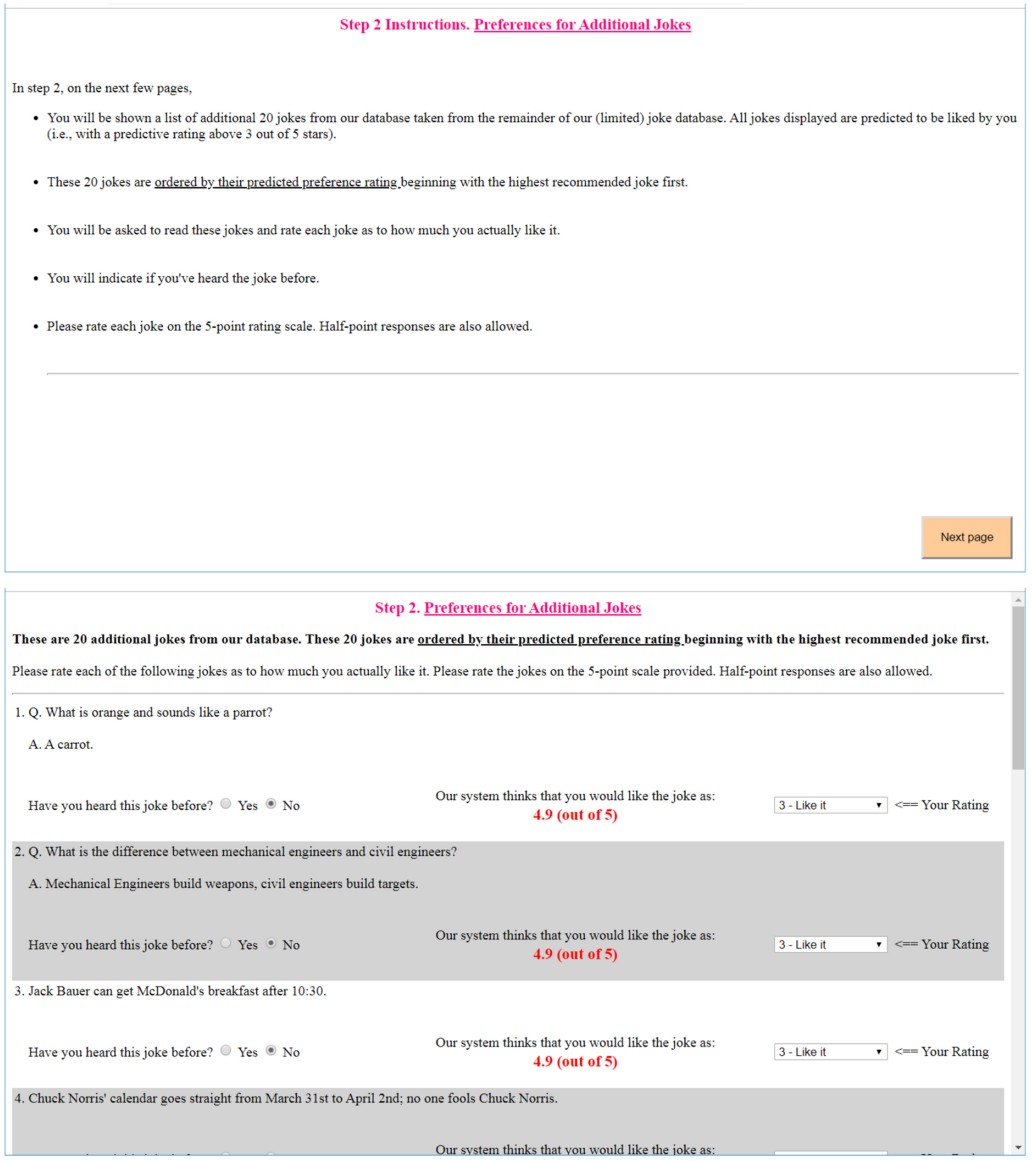


Fig. A7. Study 3: Random with predicted ratings, ordered.

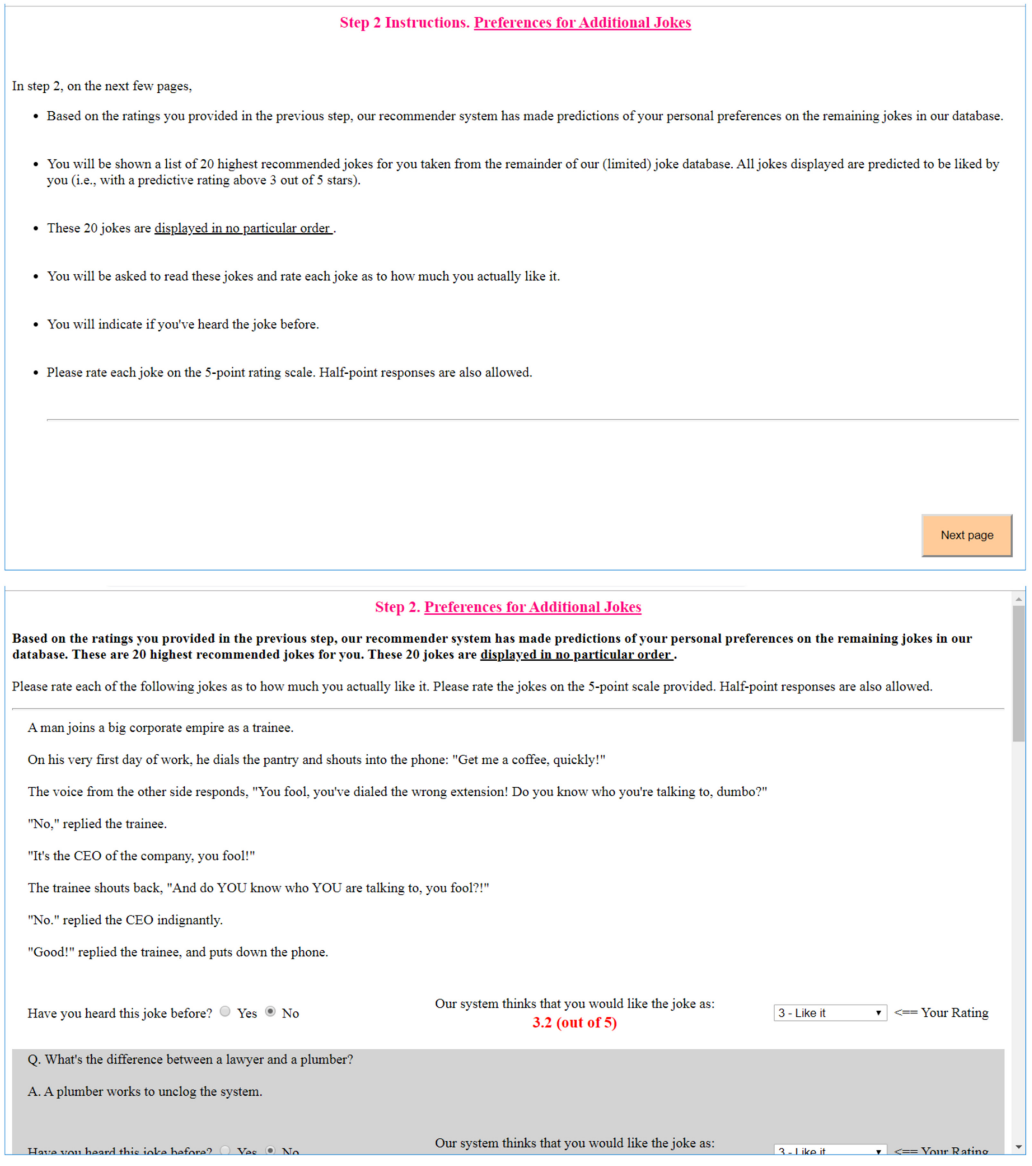


Fig. A8. Study 3: Personalized with predicted rating, unordered.

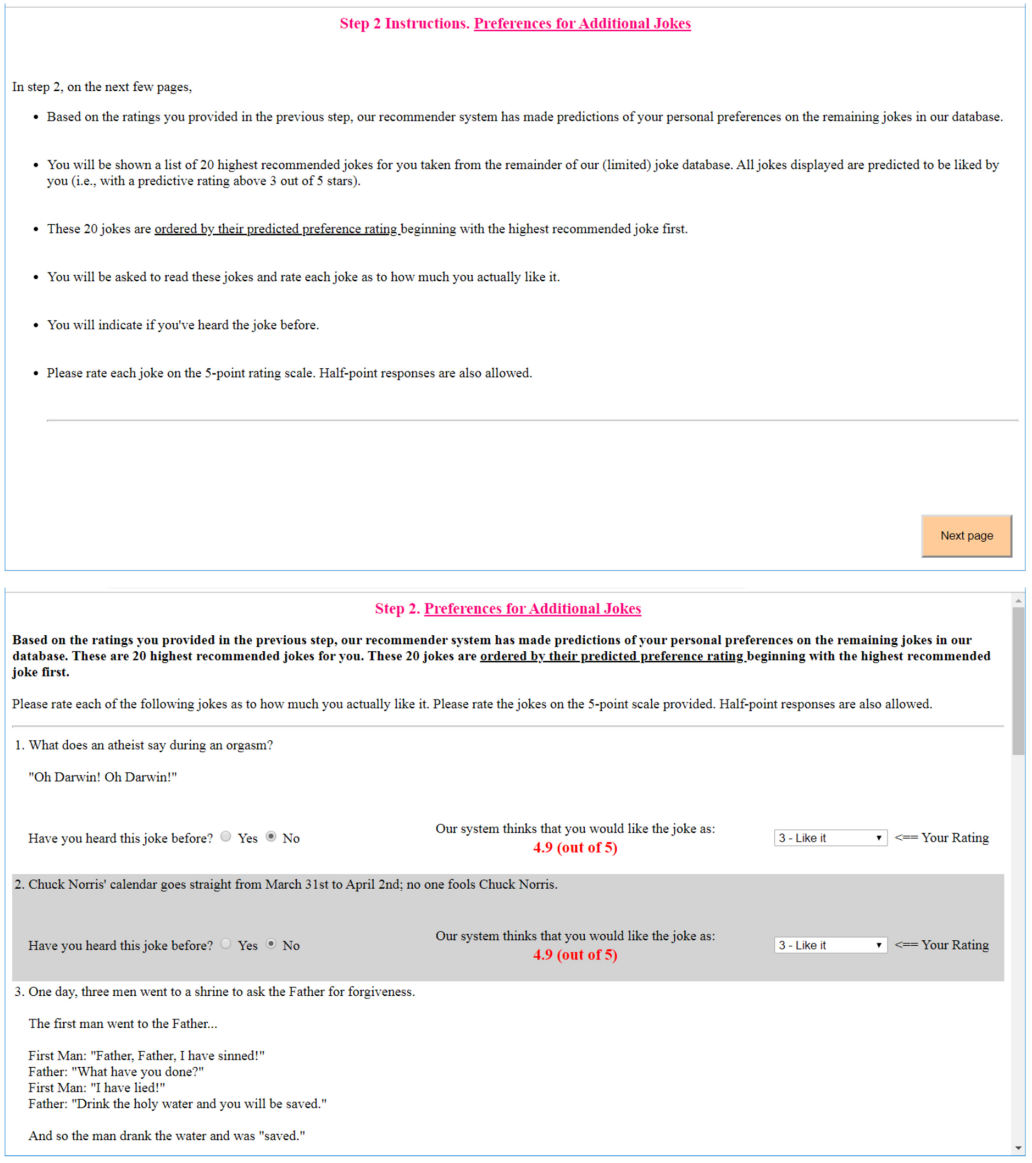


Fig. A9. Study 3: Personalized with predicted rating, ordered.

## B ALGORITHM FOR PERSONALIZED RATING PREDICTION

Using the joke rating data in the Jester dataset, we implemented a personalized recommender system that predicts participants' preferences for each joke in the database. Our system utilizes the classic item-based collaborative filtering (ItemCF) approach (e.g., Bell and Koren [2007], Breese et al. [1998], and Sarwar et al. [2001]). ItemCF is a well-known recommendation approach and is widely employed by many real-world applications, including Amazon.com [Linden et al. 2003; Smith and Linden 2017]. ItemCF makes predictions of a user's preference for each unconsumed item based on how much the user liked the target item's "nearest neighbors", i.e., other items that are similar to the target item. Specifically, for a given user $u$ and item $i$, the ItemCF approach calculates the preference rating $R^*_{ui}$ as an aggregate of user $u$'s ratings on items that have similar rating patterns to item $i$. A common aggregation approach is to use the weighted sum of the neighbors' ratings, where the similarity measure between two items is used as a weight. That is, the more similar an item $j$ and the target item $i$ are, the more weight will be carried by the rating provided on item $j$ by user $u$ in the weighted sum when computing the prediction for item $i$. Hence, the predicted rating $R^*_{ui}$ is computed as:

$$R^*_{ui} = b_{ui} + \frac{\sum_{j \in N(u,i)} sim_{ij} \times (R_{uj} - b_{uj})}{\sum_{j \in N(u,i)} |sim_{ij}|}$$

Here $N(u, i)$ is a set of "neighbors" with similar rating patterns to item $i$ that have been previously consumed by user $u$, $sim_{ij}$ is the similarity between items $i$ and $j$, and $b_{ui}$ is the baseline estimate for user $u$ on item $i$, i.e., $b_{ui} = \mu + b_u + b_i$, where $\mu$ is the overall average rating of the dataset (i.e., global effect), $b_u$ is the average observed deviation from $\mu$ on ratings provided by user $u$ (i.e., user effect), and $b_i$ is the average observed deviation from $\mu$ on ratings given to item $i$ (i.e., item effect). Normalizing the dataset by removing the baseline estimate from each rating is a suggested data pre-processing step in the literature to further improve the predictive performance of collaborative filtering recommendation algorithms [Bell and Koren 2007].

In our implementation, the similarity between two items (jokes) is calculated as the Pearson correlation coefficient between their rating vectors (based on the users who rated these items in common), as suggested by the literature [Sarwar et al. 2001]. The recommendation algorithm was tuned/pre-tested on the publicly available Jester dataset using standard machine learning practices (e.g., cross-validation). For deployment in our study, we first computed all pairwise joke similarity scores based on the Jester dataset. Then, using the joke ratings collected on the initial 50 jokes the subjects saw in our experiment, we applied the ItemCF algorithm to estimate subjects' preference ratings on the remaining jokes in the database. In all three studies, we calculated the subjects' preference ratings for jokes displayed to them in the top-N lists and controlled for their preferences in our regression analyses.

## C ROBUSTNESS CHECK WITH VERY TOP RECOMMENDATIONS WITHIN TOP-N LISTS

We conducted additional analysis to check whether looking only at top recommendations (e.g., top-2 or top-5) would give consistent results. In other words, perhaps it is possible that only top-2 or top-5 recommendations show some bias and the position of display matters only within the top few items, which then might be "diluted" if we are only looking cumulatively at the top-20 list. However, the same results hold, i.e., neither top-2 nor top-5 recommendations exhibit biasing effects, and similarly there are no item position effects within the top-2 ($b = -0.109$, $p = 0.2$) and top-5 ($b = 0.003$, $p = 0.873$) lists. The regression results are provided in Table C1 below.

Table C1. Regression Analysis of top-2 and top-5 Recommendation Lists, Study 2

| | **Top 2 Recommendations Only** | **Top 5 Recommendations Only** |
|---|---|---|
| **DV: UserRating** | **Coefficient (SE)** | **Coefficient(SE)** |
| Groups (Fully Random as base) | | |
| Ordered Aggregate Best | −0.0035 (0.189) | −0.054 (0.129) |
| Ordered Aggregate Random | − 0.167 (0.165) | − 0.068 (0.109) |
| Ordered Personalized Best | −0.357 (0.195) | −0.238 (0.138) |
| Ordered Personalized Random | −0.057 (0.167) | −0.039 (0.113) |
| PredictedRating | 0.805 (0.104)*** | 0.864 (0.081)*** |
| DisplayPosition | −0.109 (0.085) | 0.003 (0.021) |
| Controls | | |
| ifReadBefore | 0.227 (0.122) | 0.172 (0.089) |
| jokeFunniness | 0.198 (0.214) | 0.116 (0.149) |
| Age | 0.006 (0.008) | −0.003 (0.006) |
| Male | 0.146 (0.105) | 0.126 (0.074) |
| Undergrad | 0.221 (0.130) | 0.075 (0.095) |
| nativ | −0.125 (0.110) | −0.087 (0.085) |
| Constant | −0.039 (0.660) | 0.112 (0.421) |
| $R^2$ | 0.2731 | 0.2411 |
| $F(12, 194)$ | 13.87*** | 24.07*** |

***p < .001, Std. Err. adjusted for 195 clusters in userID.

## REFERENCES


H. Abdollahpouri, R. Burke, and B. Mobasher. 2017. Controlling popularity bias in learning-to-rank recommendation. In *Proceedings of the 11th ACM Conference on Recommender Systems*. ACM, pp. 42–46.

P. Adamopoulos and A. Tuzhilin. 2014. On over-specialization and concentration bias of recommendations: Probabilistic neighborhood selection in collaborative filtering systems. In *Proceedings of the 8th ACM Conference on Recommender Systems (Foster City, Silicon Valley, Calif.)*. ACM, pp. 153–160.

G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. 2013. Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inf. Syst. Res.* 24, 4 (2013), 956–975.

G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. 2016. Effects of personalized versus aggregate ratings on consumer preference responses. In *Proceedings of the Conference on Information Systems and Technology* (Nashville, TN).

G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. 2018. Effects of online recommendations on consumers' willingness to pay. *Inf. Syst. Res.* 29, 1 (2018), 84–102.

G. Adomavicius, J. Bockstedt, S. Curley, and J. Zhang. 2019. Reducing recommender systems biases: An investigation of rating display designs. *MIS Quarterly (MISQ)* 43, 4 (2019), 1321–1341.

G. Adomavicius and Y. Kwon. 2012. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Trans. Knowl. Data Eng.* 24, 5 (2012), 896–911.

G. Adomavicius and A. Tuzhilin. 2005. Toward the next generation of recommendation system: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.* 17, 6 (2005), 734–749.

A. Agarwal, K. Hosanagar, and M. D. Smith. 2011. Location, location, location: An analysis of profitability of position in online advertising markets. *J. Market Res.* 48, 6 (2011), 1057–1073.

X. Amatriain, J. M. Pujol, and N. Oliver. 2009. I like it... I like it not: Evaluating user ratings noise in recommender systems. In *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Springer, pp. 247–258.

S. Aral. 2014. The problem with online ratings. *MIT Sloan Manage. Rev.* 55, 2 (2014), 47.

R. M. Bell and Y. Koren. 2007. Improved neighborhood-based collaborative filtering. In *Proceedings of KDD Cup'07, (San Jose, Calif.)* ACM, New York, pp. 7–14.

T. Belluf, L. Xavier, and R. Giglio. 2012. Case study on the business value impact of personalized recommendations on a large online retailer. In *Proceedings of the 6th ACM Conference on Recommender Systems* (Dublin, Ireland) ACM, 277–280.

J. S. Breese, D. Heckerman, and C. Kadie. 1998. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence* (Madison, WI).

L. Chen, M. d. Gemmis, A. Felfernig, P. Lops, F. Ricci, and G. Semeraro. 2013. Human decision making and recommender systems. *ACM Trans. Interact. Intell. Syst.* 3, 3 (2013), Article 17.

D. Cosley, S. Lam, I. Albert, J. A. Konstan, and J. Riedl. 2003. Is seeing believing? how recommender interfaces affect users' opinions. In *Proceedings of the Conference on Human Factors in Computing Systems*, (Fort Lauderdale, FL) ACM New York, pp585–592.

P. Cremonesi, Y. Koren, and R. Turrin. 2010. Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the 4th ACM Conference on Recommender Systems* (Barcelona, Spain) ACM, 39–46.

M. Deshpande and G. Karypis. 2004. Item-based top-n recommendation algorithms. *ACM Trans. Inf. Syst.* 22, 1 (2004), 143–177.

J. N. Druckman and C. D. Kam. 2011. Students as experimental participants. *Cambridge Handbook of Experimental Political Science* (1), 41–57.

D. Fleder and K. Hosanagar. 2009. Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Manage. Sci.* 55, 5 (2009), 697–712.

A. Ghose, P. G. Ipeirotis, and B. Li. 2012. Designing ranking systems for hotels on travel search engines by mining user-generated and crowdsourced content. *Market. Sci.* 31, 3 (2012), 493–520.

K. Goldberg, T. Roeder, D. Gupta, and C. Perkins. 2001. Eigentaste: A constant time collaborative filtering algorithm. *Inf. Retriev.* 4, 2 (2001), 133–151.

F. Guo and D. B. Dunson. 2015. Uncovering systematic bias in ratings across categories: A Bayesian approach. In *Proceedings of the 9th ACM Conference on Recommender Systems* (Vienna, Austria) ACM, pp. 317–320.

X. He, H. Zhang, M.-Y. Kan, and T.-S. Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Pisa, Italy) ACM, 549–558.

E. Hirt, F. Kardes, and K. Markman. 2004. Activating a mental simulation mind-set through generation of alternatives: Implications for debiasing in related and unrelated domains. *J. Experi. Social Psych.* 40, 3 (2004), 374–383.

K. Hosanagar, D. Fleder, D. Lee, and A. Buja. 2014. Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Manage. Sci.* 60, 4 (2014), 805–823.

M. S. Isaac and R. M. Schindler. 2014. The top-ten effect: Consumers' subjective categorization of ranked lists. *Journal of Consumer Research* 40, 6 (2014), 1181–1202.

A. Jameson, M. C. Willemsen, A. Felfernig, M. de Gemmis, P. Lops, G. Semeraro, and L. Chen. 2015. Human decision making and recommender systems. In *Recommender Systems Handbook*, F. Ricci, L. Rokach and B. Shapira (eds.). Springer US, Boston, MA, pp. 611–648.

D. Jannach, L. Lerche, I. Kamehkhosh, and M. Jugovac. 2015. What recommenders recommend: An analysis of recommendation biases and possible countermeasures. *User Modeling and User-Adapted Interaction* 25, 5, 1, 427–491.

K. Jerath, L. Ma, Y.-H. Park, and K. Srinivasan. 2011. A position paradox" in sponsored search auctions. *Marketing Science* 30, 4 (2011), 612–627.

T. Joachims, A. Swaminathan, and T. Schnabel. 2017. Unbiased learning-to-rank with biased feedback. In *Proceedings of the 10th ACM International Conference on Web Search and Data Mining*, pp. 781–789.

S. Kabbur, X. Ning, and G. Karypis. 2013. Fism: Factored item similarity models for top-n recommender systems. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Chicago, Ill.), ACM, pp. 659–667.

Z. Kang, C. Peng, and Q. Cheng. 2016. Top-n recommender system via matrix completion. In *Proceedings of the 30th AAAI Conference on Artificial Intelligence*.

J. A. Konstan. 2004. Introduction to recommender systems: Algorithms and evaluation. *ACM Trans. Inf, Syst.* 22, 1 (2004), 1–4.

Y. Koren and R. Bell. 2015. Advances in collaborative filtering. In *Recommender Systems Handbook*, F. Ricci, L. Rokach and B. Shapira (eds.). Springer US, Boston, MA, 77–118.

S. Krishnan, J. Patel, M. J. Franklin, and K. Goldberg. 2014. A methodology for learning, analyzing, and mitigating social influence bias in recommender systems. In *Proceedings of the 8th ACM Conference on Recommender Systems*. (Foster City, Silicon Valley, Calif.) ACM, pp137–144.

G. Linden, B. Smith, and J. York. 2003. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE CS* 7, 1 (2003), 76–80.

B. Loni, R. Pagano, M. Larson, and A. Hanjalic. 2019. Top-N recommendation with multi-channel positive feedback using factorization machines. *ACM Trans. Inf. Syst.* 37, 2 (2019) Mar.

J. Lu, D. Wu, M. Mao, W. Wang, and G. Zhang. 2015. Recommender system application developments. *Decis. Support Syst.* 74, C (2015), 12–32.

W. Lu, F. L. Chung, W. H. Jiang, M. Ester, and W. Liu. 2019. A deep Bayesian tensor-based system for video recommendation. *ACM Trans. Inf. Syst.* 37, 1 (2019) Jan.

J. Monks and R. G. Ehrenberg. 1999. U.S. News & World Report's college rankings: Why they do matter. *Change: The Magazine of Higher Learning* 31, 6 (1999), 1999/11/01, 42–51.

T. Mussweiler and F. Strack. 1999. Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *Journal of Experimental Social Psychology* 35, 2 (1999), 136–164.

T. T. Nguyen, P.-M. Hui, F. M. Harper, L. Terveen, and J. A. Konstan. 2014. Exploring the filter bubble: The effect of using recommender systems on content diversity. In *Proceedings of the 23rd International Conference on World Wide Web*. (Seoul, Korea) ACM, 677–686.

X. Ning and G. Karypis. 2012. Sparse linear methods with side information for top-n recommendations. In *Proceedings of the 6th ACM Conference on Recommender Systems* (Dublin, Ireland) ACM, 155–162.

E. Pariser. 2011. *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think*. Penguin.

D. G. Pope. 2009. Reacting to rankings: Evidence from "America's Best Hospitals. *J. Health Economics* 28, 6 (2009), 1154–1165.

S. Prawesh and B. Padmanabhan. 2014. The "most popular news" recommender: count amplification and manipulation resistance. *Inf. Syst. Res* 25, 3 (2014), 569–589.

P. Resnick, R. K. Garrett, T. Kriplean, S. A. Munson, and N. J. Stroud. 2013. Bursting your (filter) bubble: Strategies for promoting diverse exposure. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work Companion* (San Antonio, TX.) Association for Computing Machinery, 95–100.

P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl. 1994. Grouplens: An open architecture for collaborative filtering of netnews. In *Proceedings of the Conference on Computer Supported Cooperative Work*, (Chapel Hill, NC), 175–186.

F. Ricci, L. Rokach, and B. Shapira. 2015. *Recommender Systems Handbook*. Springer.

Sa, #250, l. Vargas, and P. Castells. 2014. Improving sales diversity by recommending users to items. In *Proceedings of the 8th ACM Conference on Recommender Systems* (Foster City, Silicon Valley, Calif), ACM, 145–152.

B. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. 2001. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the WWW Conference* (Hong Kong). ACM, 285–295.

T. Schnabel, A. Swaminathan, A. Singh, N. Chandak, and T. Joachims. 2016. Recommendations as treatments: Debiasing learning and evaluation. *Arxiv Preprint Arxiv:1602.05352*.

S. Senecal and J. Nantel. 2004. The influence of online product recommendations on consumers' online choices. *J. Retal* 80, 2 (2004), 2004/01/01/, 159–169.

B. Smith and G. Linden. 2017. Two decades of recommender systems at amazon. com. *IEEE Internet Comput.* 21, 3 (2017), 12–18.

J. B. Soll, K. L. Milkman, and J. W. Payne. 2016. A user's guide to debiasing. In *The Wiley Blackwell Handbook of Judgment and Decision Making*. John Wiley & Sons, Ltd, 924–951.

A. T. Sorensen. 2007. Bestseller lists and product variety. *The Journal of Industrial Economics* 55, 4 (2007), 715–738.

K. Truong, F. Ishikawa, and S. Honiden. 2007. Improving accuracy of recommender system by item clustering. *IEICE Trans. on Information and Systems* (E90D:9), Sep, 1363–1373.

A. Tversky and D. Kahneman. 1974. Judgment under uncertainty: heuristics and biases. *Science* 185, 4157 (1974), 1124–1131.

A. Tversky, S. Sattath, and P. Slovic. 1988. Contingent weighting in judgement and choice. *Psychol. Review* 95, 3 (1988), 371–384.

R. Ursu. 2017. The power of rankings: Quantifying the effect of rankings on online consumer search and purchase decisions. *Market. Science.*

M. Wan, J. Ni, R. Misra, and J. McAuley. 2020. Addressing marketing bias in product recommendations. In *Proceedings of the 13th International Conference on Web Search and Data Mining*, 618–626.

N. Xia and G. Karypis. 2011. Slim: Sparse linear methods for top-n recommender systems. In *Proceedings of the 2011 IEEE 11th International Conference on Data Mining (ICDM)*, 497–506.

F. Xue, X. N. He, X. Wang, J. D. Xu, K. Liu, and R. C. Hong. 2019. Deep item-based collaborative filtering for top-n recommendation. *ACM Trans. Inf. Syst.* 37, 3 (2019) Jul.

L. Yang, Y. Cui, Y. Xuan, C. Wang, S. Belongie, and D. Estrin. 2018. Unbiased offline recommender evaluation for missing-not-at-random implicit feedback. In *Proceedings of the 12th ACM Conference on Recommender Systems* 279–287.

C. Y. Zhang, H. W. Liang, and K. Wang. 2016. Trip recommendation meets real-world constraints: Poi availability, diversity, and traveling time uncertainty. *Acm Transactions on Information Systems* 35, 1 (2016) Oct.

J. Zhang, G. Adomavicius, A. Gupta, and W. Ketter. 2019. Consumption and performance: Understanding longitudinal dynamics of recommender systems via an agent-based simulation framework. *Information Systems Research*.

T. Zhang and V. S. Iyengar. 2002. Recommender systems using linear classifiers. *Journal of Machine Learning Research* (2:Feb), 313–334.

X. Zhang, J. Zhao, and J. C. Lui. 2017. Modeling the assimilation-contrast effects in online product rating systems: Debiasing and recommendations. In *Proceedings of the 11th ACM Conference on Recommender Systems*, 98–106.

Z. Zolaktaf, R. Babanezhad, and R. Pottinger. 2018. A generic top-n recommendation framework for trading-off accuracy, novelty, and coverage. In *Proceedings of the 2018 IEEE 34th International Conference on Data Engineering (ICDE)*: IEEE, 149–160.

Received March 2020; revised August 2020; accepted October 2020