Understanding Effects of Personalized vs. Aggregate Ratings on User Preferences

Gediminas Adomavicius
University of Minnesota
Minneapolis, MN
gedas@umn.edu

Jesse Bockstedt
Emory University
Atlanta, GA
bockstedt@emory.edu

Shawn Curley
University of Minnesota
Minneapolis, MN
curley@umn.edu

Jingjing Zhang
Indiana University
Bloomington, IN
jjzhang@indiana.edu

ABSTRACT

Prior research has shown that online recommendations have significant influence on consumers' preference ratings and their economic behavior. However, research has not examined the anchoring effects of aggregate user ratings, which are also commonly displayed in online retail settings. This research compares and contrasts the anchoring biases introduced by aggregate ratings on consumers' preferences ratings to those produced by personalized recommendations. Through multiple laboratory experiments, we show that the user preferences can be affected (i.e., distorted) by the displayed average online user ratings in a similar manner as has been shown with personalized recommendations. We further compare the magnitude of anchoring biases by personalized recommendations and aggregate ratings. Our results show that when shown separately, aggregate ratings and personalized recommendations create similar effects on user preferences. When shown together, there is no cumulative increase in the effect, and personalized recommendations tend to dominate the effect on user preferences. We also test these effects using an alternative top-N presentation format. Our results here suggest that top-N lists may be an effective presentation solution that maintains key information provided by recommendations while reducing or eliminating decision biases.

Keywords

Recommender systems; personalized ratings; aggregate ratings; preference bias; anchoring effects; laboratory experiments.

1. INTRODUCTION

Most recommender systems use user ratings for previously consumed items as inputs to the system's computational techniques to estimate preferences for items that have not yet been consumed by the individual. For example, Netflix users are asked to rate the movies they have watched on a 5-star scale (with 1 being the least liked and 5 the most liked). The Netflix recommender system then analyzes patterns of users' ratings to understand users' personal interests and predict their preferences for unseen movies. Many real-world recommender systems present these estimated user preferences in the form of system-predicted ratings to indicate expectations of how much the consumer will like items, serving as recommendations. Here we use the term "recommendation" broadly to encompass any rating that is displayed to the user with the intention to convey some item "quality" information, including the ratings at the low end of the scale (such as the recommender system's predictions that the user will dislike the item). After users experience or consume the suggested item(s), they can submit their input in the form of item ratings back to the recommender system, which are used to analyze the system's accuracy and improve future recommendations, completing a feedback loop that is central to a recommender systems' use and value (Adomavicius et al. 2013).

Recent studies show that interacting with online personalization and recommendation systems can have unintended effects on user preferences and economic behavior (Cosley et al. 2003; Adomavicius et al. 2012, 2013). In particular, users' self-reported judgments can be significantly distorted by the system's predictions. 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. (2012) found that personalized ratings displayed to users significantly swayed their willingness to pay for items in the direction of the system-displayed rating value.

In addition to *personalized ratings* for products (representing estimated preferences of individual users), *aggregate ratings* for products (representing population-level preference consensus) are another important type of information on which users often rely to make their product purchase or consumption decisions. While in both cases the information often is presented in an identical or nearly identical manner (e.g., as numeric scale values or star ratings), their underlying meanings are very different. However, no prior research has systematically examined or compared the potential biases caused by information presented as aggregate information about other users' evaluations vs. personalized preference predictions from a recommender system. Therefore, we explore four issues related to the impact of aggregate vs. personalized ratings:

(1) The *bias* issue: Are users' self-reported preference ratings for products drawn toward the displayed aggregate values? In other words, do preference biases that have been observed with personalized ratings extend to non-personalized, aggregate ratings?

(2) The *relative effect size* issue: Observing which type of information results in a higher effect size: aggregate or personalized ratings?

(3) The *combination* issue: What is the combined effect of providing both aggregate and personalized ratings, as compared to receiving one alone?

For the fourth issue, we first note that, instead of displaying rating information as numeric values (e.g., personalized system-predicted ratings or mean aggregate ratings), sometimes systems use rating information to compile and display top-N item lists of "best" (or recommended) items. As examples, many news websites recommend the top-10 articles to their readers based on their interests and browsing histories, and Amazon suggests lists of products that their customers might find interesting. These recommendations are merely displayed on the webpage as a list of items (either ranked or not-ranked) often with no underlying rating values. Studying this type of presentation format provides a robustness check on the biasing effects of personalized and aggregate ratings.

(4) The *presentation format* issue: Are users' self-reported preference ratings for products still influenced by the displayed information, when the aggregate and personalized ratings are used to produce the recommendations as a list of the top-N items but without displaying explicit numeric rating values?

We conducted three controlled laboratory experiments, in which the recommendations based on aggregate and/or personalized ratings presented to participants were manipulated to answer the aforementioned research questions. In all studies, participants were asked to read a number of jokes, reporting their preference rating immediately after reading each joke.

2. BACKGROUND

Research studies in decision making, behavioral economics, and marketing have consistently observed judgmental biases across a variety of context settings (e.g., see contributions in Gilovich et al. 2002). In lay usage, the term "bias" has a negative connotation, suggesting a negative prejudice. In behavioral economic and decision research, however, the word bias is used in a more agnostic manner to represent a systematic pattern of deviation from a norm or rational standard of judgment (e.g., Haselton et al. 2005). Decision biases are not always detrimental for the decision maker; instead, the term bias highlights predictable tendencies that judgments follow under certain decision conditions.

In the context of users responding to recommendations, we define bias relative to a rational standard that is at least implicit, if not explicit, in all real-world instances of their use. The presumption is that the consumer's stated preference rating is a non-adulterated expression of their preference for the product or experience itself, as tailored to the provided scale. This should particularly be true when there is no delay between the experience or consumption of the item being rated and the reporting of the preference. This standard parallels the normative principle of invariance described and tested by Tversky and Kahneman (1986) in their discussion of framing effects as a judgmental bias. If a personalized recommendation significantly impacts the stated preference, signaling an adulteration of the preference, then we say that a decision bias has been introduced by the recommendation.

In identifying bias in this way, it is important to recognize the timing of the connection between the system's recommendation (which may include personalized or aggregate ratings for an item) and the consumer's subsequently submitted preference rating. Prior to experiencing the item, consumers seek and receive recommendations as a way of guiding their choices, purchases, and/or expectations concerning the item. At this *pre-consumption stage*, the recommendations represent a highly valuable service to consumers, providing help for finding and selecting relevant items and managing the potential information overload in many online settings. Once the item has been experienced, however, the recommendation is not assumed to provide value in the assessment of the user's inherent preference for that item. This is particularly true if the item has just been consumed, i.e., there is no potential uncertainty about the experience due to recall effects (e.g., when users are trying to remember how they felt about a movie that they saw a year ago). Recommendations are designed to provide value at the pre-consumption stage; if they impact *post-consumption* preferences as well, this represents a bias relative to the presumed standard of unpolluted preference.

Within the context of recommender systems, a few studies have explored how personalized system-predicted ratings influence online consumer behavior. These studies have shown strong and consistent evidence that ratings provided by consumers are biased toward system-generated recommendations when consumers construct their judgments for products. 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 prediction was provided. However, when users re-rated a movie while being shown a "predicted" personalized value that was altered upward or downward from their system's actual prediction by a single fixed amount of one point (i.e., providing a 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 could affect users' ratings based on preference recall, for movies seen in the past and now being evaluated. Additional recall-based effects are explored by Bollen et al. (2012).

More recently, Adomavicius et al. (2013) examined system effects in three laboratory studies for items elicited at the time of item consumption. The design removes possible explanations deriving from the preference uncertainty that can be present at the point of recall, i.e., 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.

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. 2012). 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, even when controlling for participants' preferences and demographics. The effects persisted even when item uncertainty was reduced, in this case 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 predictions and user ratings creates a demand effect to increase the correspondence between stimulus and response. In the study, willingness-to-pay judgments 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. Overall, the biases resulting from system recommendations on preference judgments have been shown to be robust across a variety of digital goods, settings, and conditions.

Further, these biases can be potentially harmful in several ways (Cosley et al. 2003; Adomavicius et al. 2013). From the consumers' perspective, recommendation biases can distort (or manipulate) their preferences and their purchasing behavior and, therefore, lead to distortions in their self-reported preference ratings and suboptimal product choices. From the firm's perspective (e.g., Amazon, Netflix), these biases may allow third-party agents to manipulate the recommender system so that it operates in their favor. This would reduce consumers' trust in the recommender system and harm its value in the long term. From the system designers' perspective, the distorted user preference ratings that are subsequently submitted as consumers' feedback can pollute the inputs of the recommender system, reducing its effectiveness.

In addition to personalized system-predicted ratings, *aggregate ratings* represent an alternate source of information about item "quality" that is directly relevant to users' decision making. In particular, they can be viewed as non-personalized (i.e., same for all users) recommendations that indicate the population-level consensus about the general quality level of a given item. This information may be derived from peer ratings or from aggregating sales, download, or click data. Similar to personalized recommender system predictions, aggregate rating information may be communicated to consumers in the form of numeric values, such as mean peer ratings, or can be used to construct top-N lists of generally best-liked items.

Note that the nature of the rating effects can be studied from both *macro-level* and *micro-level* perspectives. The market, macro-level perspective investigates the market effects of how ratings and recommendations impact sales, downloads, or other aggregate outcomes of interest to retailers. From the consumer, micro-level perspective, the interest is on how ratings and recommendations impact the behaviors of individual users.

The study of macro-level outcomes has been an active area of research investigation in recent years. For example, with respect to the effects of providing aggregate ratings, Tucker and Zhang (2011) studied the impact of popular bestseller listings based on previous clicks upon the number of future clicks received. Similarly, Godinho de Matos et al. (2016) investigated the influence of peer ratings, expressed as a list of most popular movies, on market sales within a natural field experiment. Also, using an experimental methodology by creating a music market of unknown songs and artists, Salganik et al. (2006) manipulated whether or not the participants saw the number of downloads made by others and studied the effect of this social influence on market factors. An example of academic research that investigated the market-level effects of personalized recommendations is a study by Fleder and Hosanagar (2009) who, using analytical modeling and simulation, suggested that recommendation systems can lead to a rich-get-richer effect for popular products, resulting in a decrease in sales diversity in the aggregate. Somewhat in contrast, results of Fleder and Hosanagar also suggested that personalization technologies help users to widen their interests, increasing the likelihood of commonality with others.

In contrast, the micro-level effects have been underexplored in research literature, especially with respect to the aggregate ratings and their impact on individual consumer preferences. Therefore, we focus on this issue in our current study: How do personalized system-predicted ratings and aggregate peer ratings compare and contrast as influences on individual users' reactions, particularly in the preference bias that they produce?

This comparison is particularly interesting, because the influences of personalized vs. aggregate ratings on user behavior are hypothesized along quite different psychological mechanisms. For example, supplying aggregate data within a music market, Salganik and Watts (2008) demonstrated the effect that aggregated popularity feedback had upon individual-level responses in terms of choices to listen to and download songs. The mechanism for the effect derives from social motivations, grounded in the literatures on social influence. The general dynamic is one in which the consumer engages in a form of observational learning of how to behave based on the behavior of others. In contrast, personalized recommendations do not arise from social comparison. Depending on the recommendation algorithm, the personalized system-predicted rating may or may not have any connections with others' behavior. For example, content-based algorithms depend on matching feature characteristics, not on the preferences of other users (Ricci et al. 2011). Even algorithms that incorporate preferences of other users, e.g., collaborative filtering techniques (Ricci et al. 2011), generally do not make the connection explicit or obvious to the consumer. Therefore, rather than mechanisms grounded in social psychology, the effects of personalized ratings can be posited on bases of anchoring, information integration, and processing explanations. For example, one proposed mechanism is in terms of scale compatibility, as mentioned above. Another sample mechanism proposed for the effects of personalized recommendations is an information integration explanation whereby the system-predicted rating is perceived as a piece of information that the user should use in constructing their judgment (cf. Mussweiler and Strack 1999).

3. STUDY 1: Individual Effects of Aggregate vs. Personalized Ratings

3.1 Design

This study focused on research questions (1) and (2). All the studies described in this paper involved the consumption and rating of jokes, so the participant population required no special characteristics. Participants were 118 recruits from a US college's research participant pool. Participants were paid a fixed $10 fee for completing the study. Demographic features of the sample are summarized in Table 1 separately for each of the two conditions of the between-subjects component of the design. Participant characteristics are comparable between the two treatment groups. The mean time for completing the study was 29.03 minutes, which suggests subjects invested ample time and that fatigue was not an issue.

Table 1. Demographic characteristics of participants in Study 1.

	Personalized Rating	**Aggregate Rating**
# of Participants	59	59
% Female	45.8%	47.5%
Age: Mean (SD)	23.6 (8.70)	24.0 (9.03)
% Native English Speaker	50.9%	61.0%
% Undergraduate	64.5%	55.7%

Our study used 100 jokes from the Jester joke database, which has been extensively used in prior literature (Goldberg et al. 2001; Adomavicius et al. 2013). Jokes are stimuli that can be experienced in the lab session, so that the readers' preference ratings can be gathered immediately after the reading of each joke; there is no uncertainty of preference due to memory effects. As noted in Section 2, the standard assumption in such a situation is that the user's rating should provide an unadulterated expression of the reader's preference, forming the normative expectation against which bias is defined.

The between-subjects manipulation in the study is based on the type of rating information that is presented to study participants: personalized vs. aggregate ratings. In other words, the information is presented as either a personalized rating from a recommender system or as a mean rating of other users. Each participant saw only one type of rating information, either the aggregate or personalized ratings, for all the jokes.

Using a 5-star rating scale (allowing half-star ratings), participants first evaluated 50 jokes, which were randomly selected from the list of 100 and randomly ordered. These ratings provided a guise of collecting data from which to derive personalized recommendations, and also allowed us to calculate rating predictions for use in the analysis as a control for individual differences in preference.

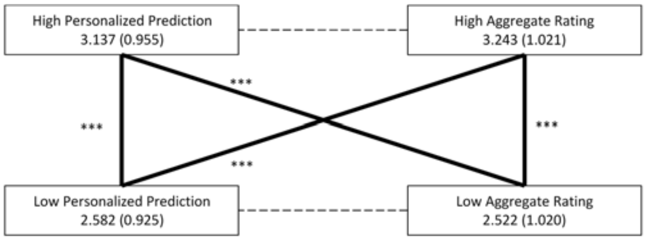
Next, the subjects received 45 jokes with rating-based information displayed. Half of the subjects randomly received the information in the form of aggregate ratings displayed as “Average user rating of this joke is: X (out of 5)”, while the other half in the form of personalized ratings displayed as “Our system thinks you would rate the joke as: X (out of 5)”. Here X is a specific rating value that was assigned separately for each joke. In both of these treatment groups (referred to as AggregateOnly and PersonalizedOnly groups), the participants saw 45 jokes in three within-subjects conditions. Specifically, 20 of these jokes were assigned to the High condition, which consisted of randomly-generated high values between 3.5 and 4.5 stars (drawn from a uniform distribution); another 20 jokes were assigned to the Low condition, which consisted of randomly-generated low values between 1.5 and 2.5 stars (drawn from a uniform distribution); and the remaining 5 jokes were assigned as the Medium condition which included randomly generated values between 2.5 and 3.5 (drawn from a uniform distribution). These 45 jokes were randomly intermixed. The Low and High conditions were oversampled since the High-Low comparison is the test of bias in this setting – i.e., whether the participants would report their post-consumption preference rating differently after being exposed to High vs. Low rating from the system. The Medium condition is included so that the presented ratings could cover the entire spectrum of the 1-5 rating scale; this helps to avoid possible credibility issues caused by bipolar recommendations (i.e., having either very high or very low ratings displayed). The responses to the Medium rating items are only useful in addressing the more peripheral issue of whether there is asymmetry in the bias between the High and Low ranges (comparing the difference of differences between High-Medium and Medium-Low), an issue that is not addressed by this paper. Hence, the Medium ratings are not analyzed here.

Finally, participants completed a short survey that collected demographic and some other individual information for use in the analyses (e.g., see Table 2 for demographics).

3.2 Results

The pairwise t-test comparisons for the High and Low treatments are illustrated in Figure 1. Both aggregate and personalized ratings generate substantial effects on post-consumption preference ratings. Users reported significantly inflated preference ratings in the high conditions as compared to the low conditions (by more than a half-star on average, overall). This result for the personalized ratings is consistent with prior results (e.g., Adomavicius et al. 2013). The result for the aggregate ratings addresses research issue (1): Even though the two types of recommendation represent very different information, they both tend to generate biases of preference ratings.

With respect to research issue (2), we note that the two Low conditions (Low personalized ratings vs. Low aggregate ratings) do not significantly differ, and that the same holds for the two High conditions. This provides evidence that aggregate and personalized ratings, when presented individually, generate similar levels of preference bias.



Note: ***p < .001, ** p < .01, [----] p > .05

Figure 1. Mean (standard deviation) of self-reported user preference ratings after observing either High vs. Low personalized or aggregate ratings. (All tests are one-tailed except for the tests represented by the horizontal lines in the figure. Unlike the others, these tests have no prior hypothesized direction, so two-tailed tests are performed.)

To get a more direct answer for research issue (2), with controls for possible confounding factors, we also employed regression analysis. Specifically, the repeated-measures design of the experiment, wherein each participant was exposed to both high and low ratings in a random fashion, allows us to model the aggregate relationship between shown ratings (either personalized or aggregate) and user’s submitted post-consumption preference ratings while controlling for individual participant differences. While we do not include the full regression analysis here due to the space limitations, the results confirm what we have observed in Figure 1 – i.e., the social, population-level information provided by aggregate ratings creates a level of bias comparable to that produced by personalized recommendations despite their different underlying mechanisms.

4. STUDY 2: Combined Effects of Both Personalized and Aggregate Ratings

4.1 Design

We recognize that, in some instances, retailers provide both pieces of information (personalized *and* aggregate ratings) as aids to the consumer. Study 2 addresses research issue (3), i.e., the issue of the *combined* effects of these two different types of ratings.

A US college’s research participant pool provided 55 participants who were paid a fixed $10 fee for completing the study. None of the participants from Study 1 were allowed to enroll in Study 2. Demographic features of the sample are summarized in Table 2 for the two groups (discussed below) of the between-subjects component of the design. Participant characteristics are comparable between the two treatment groups and to those in Study 1 drawn from the same population. The mean time for completing the study was 29.97 minutes.

Table 2. Demographic characteristics of participants in Study 2.

	Personalized Rating First	**Aggregate Rating First**
# of Participants	28	27
% Female	35.7%	48.2%
Age: Mean (SD)	24.1 (7.07)	25.7 (12.26)
% Native English Speaker	67.9%	70.4%
% Undergraduate	55.4%	56.78%

The objective of Study 2 is to examine the relative importance of personalized vs. aggregate ratings when both types of information are displayed, controlling for any order effects. Participants were randomly assigned into one of two treatment groups. Participants in both groups received both personalized and aggregate ratings for each joke displayed to them. The first group received the personalized rating first, followed by the aggregate rating (i.e., the PersonalizedFirst group). The second group saw the ratings for each joke in the reverse order (i.e., the AggregateFirst group).

The study followed a similar procedure as Study 1. Participants first read 50 randomly selected jokes from a database of 100, being asked to provide their preference for each joke using a 5-point rating scale. Each participant was then asked to rate 45 additional jokes along with both personalized and aggregate ratings displayed.

In both treatment groups, the 45 jokes were randomly assigned into five within-subjects conditions. 40 of the jokes occupied a 2×2 within-subjects design crossing High and Low values of personalized and aggregate ratings. 10 jokes were assigned to the HighP-HighA condition that consisted of high values for both personalized and aggregate ratings, 10 jokes to the LowP-LowA condition that consisted of low values for both personalized and aggregate ratings; 10 jokes to the LowP-HighA condition that consisted of low personalized and high aggregate ratings, and 10 jokes to the HighP-LowA condition that consisted of high personalized and low aggregate ratings. Similar to Study 1, all the high values are randomly generated values between 3.5 and 4.5 stars, and all the low values are randomly generated values between 1.5 and 2.5 stars, drawn from a uniform distribution. The remaining 5 jokes were assigned to the Medium condition, which included randomly generated values between 2.5 and 3.5 for both personalized and aggregate ratings. As in Study 1, the Medium condition was included simply to have a credible representation of ratings from the entire spectrum of the 1-5 rating scale; the Medium ratings are not used in any subsequent analyses in the paper. The 45 jokes were randomly intermixed.

Finally, the participants were asked to complete a survey about their demographic information and joke preferences.

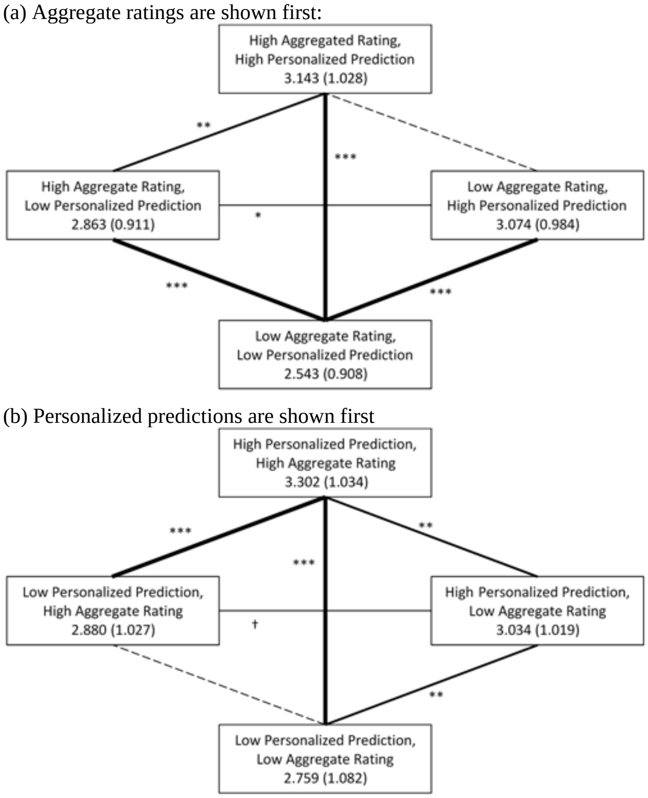
4.2 Results

Figure 2 illustrates the mean self-reported preference ratings submitted by Study 2 participants, for the cases when aggregate ratings are shown first (Figure 2a) and when personalized ratings are shown first (Figure 2b). Starting with the vertical line in each diagram, we see a strongly significant difference between when both personalized and aggregate ratings are high vs. when both ratings are low, indicating a clear bias effect in both cases, consistent with Study 1. The horizontal line in each figure indicates a stronger impact of personalized ratings compared to aggregate ratings when both appear together and signal in opposite directions (one with a high recommendation and the other low). Thus, the pairwise contrasts suggest that, when both types of ratings are present, personalized recommendations seem to be taken into account by users more strongly (and, hence, generate more significant biases) than aggregate ratings regardless of the presentation order.

This pattern is also supported by the diagonal lines in Figure 2. Beginning with the negatively sloped diagonals in the figure, i.e., when the aggregate rating goes from low to high, holding the valence of the personalized prediction fixed, the effect is variable. In each case, one of the comparisons is statistically significant, and the other is not. Specifically, a low to high difference for aggregate ratings is only observed when the aggregate ratings are preceded by high personalized predictions (Figure 2b) or followed by low personalized predictions (Figure 2a). In contrast, from the comparisons indicated by the positively sloped diagonals in the figure, i.e., when the personalized prediction goes from low to high, holding the valence of the aggregate ratings fixed, we see a clear, consistent, statistically significant biasing effect.

By comparing Figures 1 and 2, one can observe that the High vs. Low effect magnitudes are of around 0.55-0.72 when ratings are displayed individually, and around 0.55-0.6 when two ratings are displayed together. In other words, the preliminary evidence suggests that the cumulative effect of both ratings is not greater than the effect of either one of the ratings individually.

(a) Aggregate ratings are shown first:



(b) Personalized predictions are shown first

Note: ***p < .001, ** p < .01, * p < .05, † p < .10, [----] p > .10

Figure 2. Mean (standard deviation) of user preference ratings when both personalized and aggregate ratings were displayed. (All tests are one-tailed except for the tests represented by the horizontal lines in the figure. Unlike the others, these tests have no prior hypothesized direction, so two-tailed tests are performed.)

To test for the robustness of the combined effects of the two types of ratings (as compared to when each rating is shown individually) and to allow us to control for possible confounding factors, we conduct regression analyses. To do so, we pool the conditions from Studies 1 and 2 (that sampled from the same participant population) to compare the effect sizes when different amounts of information are presented to users. Recall that the High-Low comparison is the test of bias in our setting. Therefore, we include all the observations from Study 1, while for Study 2 we only consider the cases when both high personalized and high aggregate ratings (i.e., HighP-HighA) and when both low personalized and low aggregate ratings (i.e., LowP-LowA) are displayed to users. The random-effects GLS model using robust standard errors, clustered by participant, and using participant-level controls represents our model for the analysis:

$$\begin{aligned} UserRating_{ij} = {} & b_0 + b_1(High_{ij}) + b_2(Group_i) + b_3(High_{ij}*Group_i) + \\ & b_4(Deviation_{ij}) + b_5(PredictedRating_{ij}) + \\ & b_6(AdditionalControls_{ij}) + u_i + \varepsilon_{ij} \end{aligned}$$

In the model, $\mathit{UserRating}_{ij}$ is the submitted rating for participant *i* on joke *j*. High_{ij} is a binary variable that indicates whether the shown rating for participant *i* on joke *j* is a high or low artificial rating. Thus, the coefficient on High_{ij} measures the difference in user preference ratings between high and low conditions (manipulated within-subjects), which is our operationalization of bias. To the extent that users are influenced by the observed information, their submitted preference ratings will be shifted up when seeing High ratings and shifted down when seeing Low ratings. Thus, the High/Low difference is an indicator of the bias created by the observed information. $\mathit{AggregateDev}_{ij}$ and $\mathit{PersonalizedDev}_{ij}$ are derived variables that capture the deviations between the actually shown rating (aggregate and personalized, respectively) for participant *i* on joke *j* and the expected value for the shown ratings in the corresponding condition. Recall that the two ratings seen by the participant were manipulated independently to introduce randomness into the values for the high and low rating conditions by drawing from uniform distributions: High [3.5, 4.5] and Low [1.5, 2.5]. Thus, the deviation variables are computed by either subtracting 4.0 from the shown rating (High condition) or 2.0 from the shown rating (Low condition). Group_i denotes different between-subject conditions with different information displays. For Model 1, the groups are *AggregateOnly* (from Study 1) and the two order conditions from Study 2: *AggregateFirst* and *PersonalizedFirst*. For Model 2, the AggregateOnly condition is replaced by the *PersonalizedOnly* condition. We also include the interaction term between Group_i with the rating value (i.e., $\mathit{High}_{ij}*\mathit{Group}_i$). The interaction term examines whether the effect size of showing high vs. low information differs among groups. $\mathit{PredictedRating}_{ij}$ is the system-predicted rating for participant *i* on joke *j*, calculated after the fact.[1] The inclusion of this term provides an important control for differing expected joke preferences among participants. The collection of the ratings on the first 50 jokes in the procedure (without recommendations) allowed us to calculate these predicted ratings for all subjects. $\mathit{AdditionalControls}_{ij}$ is a vector of joke and participant-related variables. The controls included in the model were: the joke's funniness (average joke rating in the Jester dataset, continuous between 0 and 5), participant age (integer), gender (binary), school level (undergrad yes/no binary), whether they are native speakers of English (yes/no binary), whether they thought recommendations in the study were accurate (interval five-point scale), whether they thought the recommendations in the study were useful (interval five-point scale), and whether they thought that recommendations in general were useful (interval five-point scale).

The study utilized a repeated-measures design with a balanced number of observations on each participant. To control for participant-level heterogeneity, the composite error term ($u_i + \varepsilon_{ij}$) includes the individual participant effect u_i and the standard disturbance term ε_{ij}. A random-effects model is used for participant heterogeneity, since these individual-specific effects are uncorrelated with the randomly applied treatment conditions.

Table 3 summarizes our pooled regression analyses. The models are analyzed separately to allow us to use a full set of control variables for each regression model. Model 1 includes the conditions in which aggregate ratings are displayed, i.e., AggregateOnly from Study 1, and the HighP-HighA and LowP-LowA conditions from Study 2. Model 2 includes the conditions in which personalized predictions are displayed, i.e., PersonalizedOnly from Study 1 and the HighP-HighA and LowP-LowA conditions from Study 2.

Interestingly, in both models, the interaction term $\mathit{High}_{ij}*\mathit{Group}_i$ is statistically insignificant. There is no evidence that the effect size measured in high-low difference differs among groups with different information displays. In other words, when aggregate and personalized ratings are presented together (as explored in Study 2) regardless of their order, there is no evidence that they increase preference biases compared to displaying either aggregate or personalized recommendations alone. In other words, our results suggest a substitutionary effect between personalized and aggregate ratings on users' self-reported preferences (so that the cumulative effect is no greater than the effect of only one).

Table 3. Pooled Regression analyses for Studies 1 and 2.

DV: UserRating	Model 1: AggrOnly (Study1), All Data (Study2)	Model 2: PersOnly (Study1), All Data (Study2)
	Coefficient (SE)	Coefficient (SE)
High	0.718(0.06)***	0.551(0.068)***
Group		
(Baseline: AggregateOnly or PersonalizedOnly)		
AggregateFirst	-0.037(0.1)	-0.06(0.091)
PersonalizedFirst	0.118(0.103)	0.109(0.097)
High * Group		
High * AggregateFirst	-0.078(0.138)	0.074(0.142)
High * PersonalizedFirst	-0.135(0.138)	0.018(0.142)
AggregateDev	0.249(0.048)***	
PersonalizedDev		0.264(0.055)***
PredictedRating	0.838(0.073)***	0.679(0.054)***
Control		
jokeFunniness	0.163(0.105)	0.182(0.088)*
Age	-0.003(0.004)	0(0.002)
Male	0.006(0.067)	0.045(0.048)
Undergrad	0.09(0.077)	0.025(0.055)
Native	0.001(0.065)	-0.133(0.046)*
AggregateAccurate	-0.032(0.028)	
AggregateUseful	0.009(0.022)	
GeneralAggregateUseful	0.011(0.014)	
PersonalizedAccurate		-0.07(0.025)*
PersonalizedUseful		0.067(0.021)***
GeneralPersonalizedUseful		-0.002(0.015)
Constant	-0.368(0.301)	0.107(0.31)
N	114	114
R^2 within-subject	.2204	.1696
R^2 between-subject	.6148	.6605
R^2 overall	.3024	.2474
χ^2	992 (15 df), p < .0001	676 (15 df), p < .0001

5. STUDY 3: Recommendation List Effects

5.1 Design

In Studies 1 and 2, the recommendations were presented to users as values along a 1-5 scale. As noted earlier, another common format for providing recommendations is in the form of lists of recommended items. The numeric values corresponding to the list items often are not displayed, even when these recommendations are generated by selecting items based on the personalized rating

[1] We applied the well-known item-based collaborative filtering (CF) technique (Deshpande and Karypis 2004; Sarwar et al. 2001) to implement a recommender system that estimated users' preference ratings for the jokes. Item-based CF is one of the most popular techniques used in real-world applications because of its efficiency and accuracy. This technique allows us to precompute the main portion of our recommendation model (i.e., the similarity scores between items based on their rating patterns) in advance based on the extensive Jester rating dataset.

predictions or aggregate peer rating values. In Study 3, we examine whether the recommendations (based on either aggregate or personalized ratings) presented as a list of top-N items can lead to bias in users' reported preference ratings. Doing so explores the generalizability of the results of Studies 1 and 2, addressing research issue (4): As a non-numerical format (derived from numerical information), when compared to explicitly provided numeric ratings, do top-N recommendation lists create bias in users' self-reported post-consumption preference ratings?

Recruited from the same population as for Studies 1 and 2, 184 new participants were paid a fixed $10 fee for completing the study. Table 4 shows demographic features of the sample across all conditions of the between-subjects component of the design. Participant characteristics are comparable to those of Studies 1 and 2. The mean time for completing the study was 20.13 minutes.

Table 4. Demographic characteristics of participants in Study 3.

# of Participants	184
% Female	58.7%
Age: Mean (SD)	22.8 (7.70)
% Native English Speaker	76.6%
% Undergraduate	77.7%

The procedure followed that used in Studies 1 and 2. Participants first went through a list of 50 randomly selected jokes from the 100 jokes in the database, providing ratings using the 5-point rating scale. Participants then saw 20 additional jokes displayed as a list, and they rated the jokes using the same 5-point rating scale. Finally, the participants were asked to complete a survey about demographic information and joke preferences.

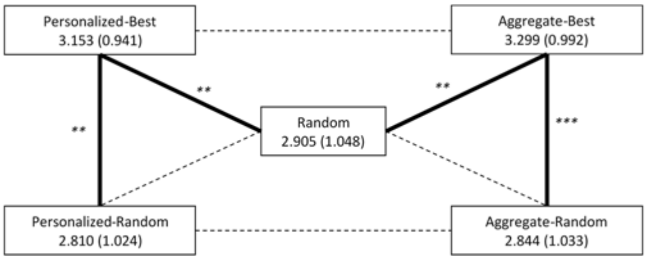
Table 5. Experimental conditions used in Study 3.

Group	N	System Description	Actual Operationalization
Random	34	"These are 20 additional jokes from our database (in no particular order)."	A list of 20 randomly-selected jokes were displayed to all participants.
Personalized-Random	34	"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.
Personalized-Best	39		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.
Aggregate-Random	34	"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.
Aggregate-Best	43		A list of top 20 jokes with the highest mean user ratings were selected.

Each participant was randomly assigned into one of five treatment groups: Random, Personalized-Random, Personalized-Best, Aggregate-Random, and Aggregate-Best. Table 5 summarizes the five conditions and the number of respondents in each condition. In the Random condition, participants were told that the additional 20 jokes were randomly selected from an existing database and displayed in no particular order, and these jokes were indeed random selections from the database in the actual operationalization. For the two personalized conditions, both groups of participants were told that the item list contains the top jokes selected by a recommender system based on their preferences. The difference between Personalized-Random and Personalized-Best lies in the actual operationalization of selecting the jokes. In the Personalized-Random condition, the jokes displayed to the participants were actually randomly selected jokes from the database. In the Personalized-Best condition, the jokes were the *actual* top 20 jokes that had the highest predicted ratings estimated by the well-known item-based collaborative filtering recommendation algorithm (Deshpande and Karypis 2004; Sarwar et al. 2001). Similarly, for the two aggregate conditions, both groups of participants were told that the item list contains the top jokes selected according to aggregate user ratings on these jokes. In the Aggregate-Random condition, the jokes displayed to the participants were actually randomly selected jokes from the database. In the Aggregate-Best condition, the jokes were the actual top 20 jokes that had the highest mean user ratings based on the Jester dataset. To control for joke funniness, we displayed the same list of 20 jokes to all participants in the Personalized-Random, Aggregate-Random and Random conditions, albeit the display order was shuffled for each person.

5.2 Results

Figure 3 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 Random group. In other words, the top-N information presentation format did not introduce bias in users' submitted ratings. Figure 3 also suggests that the consumers were not just generally insensitive. On average, participants who received actual top-N lists, either based on personalized (Personalized-Best) or aggregate 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, since the jokes displayed in the Personalized-Best and Aggregate-Best conditions were likely better jokes for the participants.



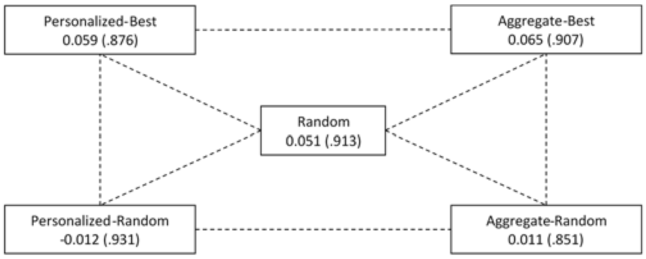
Note: ***p < .001, ** p < .01, [----] p > .05

Figure 3. Contrasts of mean preference ratings by experimental condition. (All tests are two-tailed.)

In order to separate item quality from potential bias, we conducted two further analyses. First, we adjusted the preference ratings for the viewed jokes using a rating drift measure (Adomavicius et al. 2013) defined as:

Rating Drift = Submitted Rating – System-Predicted Rating.

The system- predicted rating represents the rating of a user-joke combination as predicted by the recommendation algorithm (an item-based collaborative filtering method, in our case). Submitted rating is the user's reported rating after reading the joke. So, positive/negative rating drift values represent situations where the user's submitted rating was higher/lower than the system-predicted rating. In that the predicted rating captures a valid indicator of user's preferences based on their initial joke responses (which has been demonstrated in prior work, e.g., by Adomavicius et al. (2013)), rating drift is a measure that removes a component of individual preference from the user rating, leaving a measure that is more representative of possible bias (though still not pure, since the predicted ratings are not 100% accurate). Still, the contrast t-tests of rating drift across treatments in Study 3 are highly illuminative, as illustrated by Figure 4. As observed, rating drifts for all five experimental groups had small values that ranged from -0.012 to 0.065, suggesting that user-submitted ratings were statistically indistinguishable from system-predicted ratings. Also, the differences in rating drifts between Random and any treatment group were insignificant, suggesting that item lists presented in different ways (i.e., based on personalized recommendations, aggregate user ratings, or random lists) did not pull users away from their preferences as captured by the system-predicted ratings.



Note: [----] p > .05

Figure 4. Contrasts of mean rating drifts by experimental condition. (All tests are two-tailed.)

6. CONCLUSIONS

Prior research has shown that system recommendations can impact users' self-reported preference ratings, which can have deleterious effects (Cosley et al. 2003; Adomavicius et al. 2013). We extend this stream of research to investigate the decision biases introduced by aggregate peer ratings on users' post-consumption preference ratings. Through laboratory experiments, we first demonstrate that the self-reported preference rating (for a specific consumed item) can be strongly biased not only by observing personalized, system-predicted ratings, but also non-personalized, aggregate user ratings. We further demonstrate that, when personalized and aggregate ratings are displayed together, there is no cumulative increase in effect and that users tend to focus more attention on personalized ratings. Finally, we show that alternative recommendation displays that use top-N lists instead of individual item rating information seem to greatly reduce, if not remove, the biases observed in prior studies – in other words, this format appears to be a promising alternative for recommending items to users without introducing decision biases. This may be because top-N lists do not include explicit values (i.e., the individual system-predicted ratings or aggregate user ratings), which are likely causing the biases observed in prior studies. These results provide several obvious practical implications for the design of recommender system and online retail interfaces and displays.

7. REFERENCES

[1] Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2012. Effects of Online Recommendations on Consumers' Willingness to Pay. *ACM RecSys 2012 Workshop on Human Decision Making in Recommender Systems*, Dublin, Ireland.

[2] Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2013. Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. *Information Systems Research* (24:4), pp. 956-975.

[3] Bollen, D., Graus, M., and Willemsen, M.C. 2012. Remembering the Stars? Effect of Time on Preference Retrieval from Memory. In *Proceedings of the Sixth ACM Conference on Recommender Systems* (RecSys 2012). ACM, New York, NY, USA, 217-220.

[4] Cosley, D., Lam, S., Albert, I., Konstan, J.A., and Riedl, J. 2003. Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (CHI 2003), Fort Lauderdale, FL, pp. 585-592.

[5] Deshpande, M., and Karypis, G. 2004. Item-Based Top-N Recommendation Algorithms. *ACM Transactions on Information Systems* (22:1), pp. 143-177.

[6] Fleder, D., and Hosanagar, K. 2009. Blockbuster Culture's Next Rise or Fall: The Impact of Recommender Systems on Sales Diversity. *Management Science* (55:5), pp. 697-712.

[7] Gilovich, T., Griffin, D., and Kahneman, D. 2002. *Heuristics and Biases: The Psychology of Intuitive Judgment* (1 ed.), Cambridge University Press.

[8] Godinho de Matos M., Ferreira P., Smith M.D. and Telang R. 2016. Culling the herd: Using real-world randomized experiments to measure social bias with known costly goods. *Management Science*.

[9] Goldberg, K., Roeder, T., Gupta, D., and Perkins, C. 2001. Eigentaste: A Constant Time Collaborative Filtering Algorithm. *Information Retrieval* (4:2), pp. 133-151.

[10] Haselton, M.G., Nettle, D., and Andrews, P.W. 2005. The Evolution of Cognitive Bias. In *The Handbook of Evolutionary Psychology*, D.M. Buss (ed.), Hoboken NJ: John Wiley & Sons, pp. 724–746.

[11] Mussweiler, T., and Strack, F. 1999. Hypothesis-Consistent Testing and Semantic Priming in the Anchoring Paradigm: A Selective Accessibility Model. *Journal of Experimental Social Psychology* (35:2), 3//, pp. 136-164.

[12] Ricci, F., Rokach, L., Shapira, B., and Kantor, P. (eds.) *Recommender Systems Handbook*, Springer, 2011.

[13] Salganik, M.J., Dodds, P.S., and Watts, D.J. 2006. Experimental Study of Inequality and Unpredictability in an Artificial Cultural Market. *Science* (311:5762), pp. 854-856.

[14] Salganik, M.J., and Watts, D.J. 2008. Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market. *Social Psychology Quarterly* (74:4), Fall, p. 338.

[15] Sarwar, B., Karypis, G., Konstan, J.A., and Riedl, J. 2001. Item-Based Collaborative Filtering Recommendation Algorithms. *10th International WWW Conference*, Hong Kong, pp. 285-295.

[16] Tucker, C., and Zhang, J. 2011. How Does Popularity Information Affect Choices? A Field Experiment. *Management Science* (57:5), pp. 828-842.

[17] Tversky, A., and Kahneman, D. 1986. Rational Choice and the Framing of Decisions. In *Rational Choice: The Contrast between Economics and Psychology*, R.M.Hogarth and M.W. Reder (eds.). Chicago: Univ. of Chicago Press, pp. 67-94.