De-Biasing User Preference Ratings in Recommender Systems

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

Jesse Bockstedt
University of Arizona
Tucson, AZ
bockstedt@email.arizona.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 users' preference ratings and economic behavior. Specifically, the self-reported preference rating (for a specific consumed item) that is submitted by a user to a recommender system can be affected (i.e., distorted) by the previously observed system's recommendation. As a result, anchoring (or anchoring-like) biases reflected in user ratings not only provide a distorted view of user preferences but also contaminate inputs of recommender systems, leading to decreased quality of future recommendations. This research explores two approaches to removing anchoring biases from self-reported consumer ratings. The first proposed approach is based on a computational post-hoc de-biasing algorithm that systematically adjusts the user-submitted ratings that are known to be biased. The second approach is a user-interface-driven solution that tries to minimize anchoring biases at rating collection time. Our empirical investigation explicitly demonstrates the impact of biased vs. unbiased ratings on recommender systems' predictive performance. It also indicates that the post-hoc algorithmic de-biasing approach is very problematic, most likely due to the fact that the anchoring effects can manifest themselves very differently for different users and items. This further emphasizes the importance of proactively avoiding anchoring biases at the time of rating collection. Further, through laboratory experiments, we demonstrate that certain interface designs of recommender systems are more advantageous than others in effectively reducing anchoring biases.

Keywords

Recommender systems, anchoring effects, rating de-biasing

1. INTRODUCTION

Recommender systems are prevalent decision aids in the electronic marketplace, and online recommendations significantly impact the decision-making process of many consumers. Recent studies show that online recommendations can manipulate not only consumers' preference ratings but also their willingness to pay for products [1,2]. For example, using multiple experiments with TV shows, jokes and songs, prior studies found evidence that a recommendation provided by an online system serves as an anchor when consumers form their preference for products, even at the time of consumption [1]. Furthermore, using the system-predicted ratings as a starting point and biasing them (by perturbing them up or down) to varying degrees, this anchoring effect was observed to be continuous, with the magnitude proportional to the size of the perturbation of the recommendation in both positive and negative directions – about 0.35-star effect for each 1-star perturbation on average across all users and items [1]. Additionally, research found that recommendations displayed to participants significantly pulled their willingness to pay for items in the direction of the recommendation, even when controlling for participants' preferences and demographics [2].

Based on these previous studies, we know that users' preference ratings can be significantly distorted by the system-predicted ratings that are displayed to users. Such distorted preference ratings are subsequently submitted as users' feedback to recommender systems, which can potentially lead to a biased view of consumer preferences and several potential problems [1,5]: (i) biases can contaminate the recommender system's inputs, weakening the system's ability to provide high-quality recommendations in subsequent iterations; (ii) biases can artificially pull consumers' preferences towards displayed system recommendations, providing a distorted view of the system's performance; (iii) biases can lead to a distorted view of items from the users' perspectives. Thus, when using recommender systems, anchoring biases can be harmful to system's use and value, and the removal of anchoring biases from consumer ratings constitutes an important and highly practical research problem.

In this research, we focus on the problem of "de-biasing" self-reported consumer preference ratings for consumed items. We first empirically demonstrate that the use of unbiased preference ratings as inputs indeed leads to higher predictive accuracy of recommendation algorithms than the use of biased preference ratings. We then propose and investigate two possible approaches to tackle the rating de-biasing problem:

1) Post-hoc rating adjustment (reactive approach): a computational approach that attempts to adjust the user-submitted ratings by taking into account the system recommendation observed by the user.
2) Bias-aware interface design for rating collection (proactive approach): a design-based approach that employs a user interface for rating collection by presenting recommendations in a way that eliminates or reduces anchoring effects.

2. BACKGROUND

Prior literature has investigated how the cues provided by recommender systems influence online consumer behavior. For example, Cosley et al. (2003) found that users showed high test-retest consistency when being asked to re-rate a movie with no prediction provided [5]. However, when users were asked to re-rate a movie while being shown a "predicted" rating that was altered upward or downward from their original rating by a single fixed amount of one rating point (i.e., providing a high or low anchor), users tended to give higher or lower ratings, respectively, as compared to a control group receiving accurate original ratings. This showed that anchoring could affect users' ratings based on preference recall, for movies seen in the past and now being evaluated.

Adomavicius et al. (2013) looked at a similar effect in an even more controlled setting, in which the consumer preference ratings for items were elicited at the time of item consumption [1]. Even without a delay between consumption and elicited preference, anchoring effects were observed. The displayed predicted ratings, when perturbed to be higher or lower, affected the submitted consumer ratings to move in the same direction.

Prior research also found that recommendations not only significantly affect consumers' preference ratings but also their economic behavior [2]. Researchers present the results of two controlled experiments in the context of purchasing digital songs. The studies found strong evidence that randomly assigned song recommendations affected participants' willingness to pay, even when controlling for participants' preferences and demographics. Similar effects on willingness to pay were also observed when participants viewed actual system-generated recommendations that were intentionally perturbed up or down (introducing recommendation error).

The anchoring biases occurring due to system-generated recommendations can potentially lead to several issues. From the consumers' perspective, anchoring biases can distort (or manipulate) consumers' preferences and economic behavior, and therefore lead to suboptimal product choices and distorted preference ratings. From the retailer's perspective (e.g., Amazon, eBay), anchoring biases may allow third-party agents to manipulate the recommender system (e.g., by strategically adding malicious ratings) so that it operates in their favor. This would reduce consumers' trust in the recommender system and harm the success of the system in the long term. From the system designers' perspective, the distorted user preference ratings that are subsequently submitted as consumers' feedback to recommender systems can contaminate the inputs of the recommender system, reducing its effectiveness. Therefore, removing the bias of recommendations represents an important research question. In the following sections, we empirically study two possible approaches for tackling the rating de-biasing problem.

3. APPROACH I: POST-HOC RATING ADJUSTMENT

3.1 Rating Adjustment Algorithm

The underlying intuition of post-hoc rating adjustment is to "reverse-engineer" consumers' true non-biased ratings from the user-submitted ratings and the displayed system recommendations (that were observed by the users). For this, we use the information established by previous research that, in aggregate, the anchoring effect of online recommendations is linear and proportional to the size of the recommendation perturbation [1]. As depicted in Fig 1, the deviation of the submitted rating from the user's unbiased rating (i.e., *Dev*) should be proportional to the deviation of the system's displayed prediction from the user's unbiased rating (i.e., $\alpha \times Dev$). Given the user's submitted rating, the displayed system prediction, and the expected anchoring effect size, we develop a computational rule to systematically reverse-engineer user's unbiased ratings.

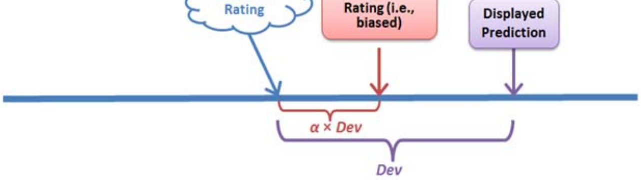


Fig 1. Post-Hoc Rating Adjustment Illustration

Mathematically, let α be the expected slope (i.e., proportionality coefficient) of the bias relative to the size of rating perturbation, R_{ui}^{Shown} be the value of the system's predicted rating on item i that was shown to user u, and $R_{ui}^{Submitted}$ be the user's submitted rating after seeing the system's prediction. We estimate the unbiased rating of user u for item i, i.e., $R_{ui}^{UnbiasedRating}$ using the formula below:

$$R_{ui}^{UnbiasedRating} = (R_{ui}^{Submitted} - \alpha \times R_{ui}^{Shown})/(1-\alpha).$$

In this post-hoc adjustment approach, the value of α is determined by the observed slope of the bias and can range between 0 (inclusive) and 1 (exclusive). Varying the size of α within [0, 1) changes the degree of rating adjustment, i.e., a larger value of α leads to a larger adjustment to the submitted rating, while $\alpha = 0$ means no adjustment is made. In our experiments, the slope α can be either a global constant that applies to all users and items, or user-specific values determined by an individual user's tendency of anchoring on the system's recommendations.

3.2 Computational Experiments

3.2.1 Joke Rating Dataset

Our experiments use a Joke rating dataset collected in laboratory settings by a prior study on anchoring effects of recommender systems [1]. The dataset includes ratings provided by 61 users on 100 jokes. At the beginning of the study, participants first evaluated 50 jokes without seeing a system's recommendations. These initial ratings reflect user's unbiased preferences and were used as a basis for computing the system's predictions. Next, the participants received 40 jokes with a predicted rating displayed. Among them, thirty of these predicted ratings were perturbed to various degrees and ten were not perturbed. These 40 jokes were randomly intermixed.

Prior research has observed continuous and linear anchoring effects on this joke rating dataset. On average, the anchoring slope across all users and items is $\alpha = 0.35$, and is significantly positive. Individual linear regression models were also obtained at an individual-user level. These user-specific regression slopes are predominately positive, suggesting that significant anchoring bias was observed for most participants.

For the post-hoc de-biasing experiments, we partition the joke ratings for each user into two subsets. The first subset contains the initial 50 ratings provided by each user before seeing any system recommendations (i.e., unbiased), and the second subset contains the subsequent 40 user ratings submitted after user received system's recommendations with various levels of perturbations (i.e., biased ratings). Next, on the 40 biased ratings, we apply the post-hoc rating adjustment rule to remove possible anchoring biases to recover users' unbiased ratings.

To evaluate the benefits of post-hoc rating adjustment, we compute predictive accuracy (measured as Root Mean Squared Error, i.e., RMSE) of standard recommendation algorithms using the adjusted ratings (i.e., de-biased) as training data and the initial ratings (i.e., unbiased) as testing data. We then compare this accuracy performance with that of using actual submitted ratings (i.e., biased) as training data and the same initial ratings as testing data. If rating de-biasing is successful, the prediction accuracy on "de-biased" ratings should be better than accuracy on "biased" ratings. We explore the post-hoc rating adjustment under a variety of settings, as described below.

3.2.2 Experiments

Our first experiment investigated the accuracy performance on unbiased, biased, and de-biased ratings adjusted based on various rules and statistically compared their differences. First, we randomly divided the 50 initial (unbiased) ratings provided by each user into two equal subsets with 25 ratings per user (aggregated across all users) in each subset. We used one subset as the training data to build the model and evaluated the model's predictive accuracy on the other subset (i.e., the testing set). Because both training and testing data are comprised of unbiased ratings submitted by users without seeing any system prediction, the accuracy performance computed based on these initial ratings would provide us the upper bound of accuracy performance for each recommendation algorithm.

We then selected 25 random ratings from the set of 40 biased submissions for each user and used them as inputs to re-build the recommendation model. The model's predictive accuracy was evaluated on the same exact testing set (i.e., 25 unbiased ratings from each user). Next we adjusted these 25 biased ratings using either the suggested global slope of α = 0.35 or user-specific adjustment slopes. When a global adjustment is used, the ratings submitted by all users are adjusted using the same global slope α. In contrast, when a user-specific adjustment is used, we first estimate the regression slope α_u for each user u based on the user's experimental data. If the estimated slope α_u is significant (i.e., p <= 0.05), we use α_u to adjust the ratings provided by the given user. Each user hence has a unique adjustment slope. Finally, we computed the predictive accuracy using these 25 de-biased ratings as training data. The predictive accuracy of rating samples was computed for several well-known recommendation algorithms, including a simple global baseline heuristic (i.e., Baseline) [3], the matrix factorization approach (i.e., SVD) [8], and user- and item-based collaborative filtering algorithms (i.e., CF_User and CF_Item) [7,10].

In our experiment we repeated the above steps 30 times and extracted different random samples each time. We report the average accuracy performances based on unbiased, biased, and de-biased ratings in Table 1. The training data resulting in best performance for each recommendation method is indicated in boldface.

Table 1. Mean predictive accuracy performance (measured in RMSE) based on different training ratings

Method	Initial (Unbiased) Ratings	Biased Ratings	De-Biased (Global Slope 0.35)	De-Biased (User-Specific Slopes)
SVD	**0.9572**	0.9663	0.9955	0.9945
CF_Item	**0.9749**	0.9968	1.0450	1.0421
CF_User	**0.9810**	1.0025	1.0568	1.0536
Baseline	**0.9521**	0.9707	1.0048	1.0046

As seen in Table 1, the initial (unbiased) ratings provide the best accuracy performance for all recommendation algorithms, clearly demonstrating the advantage of unbiased ratings over biased ratings on recommender systems' predictive performance. Most of the accuracy comparisons in the table are statistically significant (p < 0.05). The only two exceptions are the contrasts between de-biased ratings based on global and user-specific slopes for Baseline and SVD. The results suggest that the use of unbiased preference ratings as inputs indeed leads to significantly higher predictive accuracy of recommendation algorithms than the use of biased preference ratings. In addition, the de-biased ratings (adjusted based on either global or user-specific slopes) did not provide accuracy benefits. Adjusted ratings based on user-specific slopes lead to slightly better accuracy than ratings adjusted based on the global slope of α = 0.35. However, neither of the two post-hoc de-biasing adjustments was helpful in improving accuracy. These patterns are consistent across various popular recommendation algorithms described in Table 1.

In the second experiment, we explored different de-biasing slope values for user ratings and computed predictive accuracy on the entire rating dataset (as opposed to randomly chosen rating samples as in first experiment). Specifically, we took all 40 biased ratings submitted by users after seeing the system's predictions and adjusted these ratings using the post-hoc de-biasing rule. All of these 40 "de-biased" ratings were then used as training data to compute predictions using standard recommendation algorithms, and the predictive accuracy was evaluated on the initial 50 unbiased ratings. We varied the de-biasing slopes and explored both global and user-specific adjustments.

Fig 2 summarizes the predictive accuracy performance on ratings de-biased based on different adjustment slope parameters. When the slope value is equal to zero, it means no adjustment was made, i.e., the user's actual submitted ratings (biased) were used as training data for the recommendation algorithms. The vertical black line on the left side corresponds to the accuracy performance of various algorithms with these actual-submitted ratings (i.e., biased) as training data. In addition to exploring different global adjustment slopes, we also experimented with user-specific adjustments as indicated by the vertical black line on the right side.

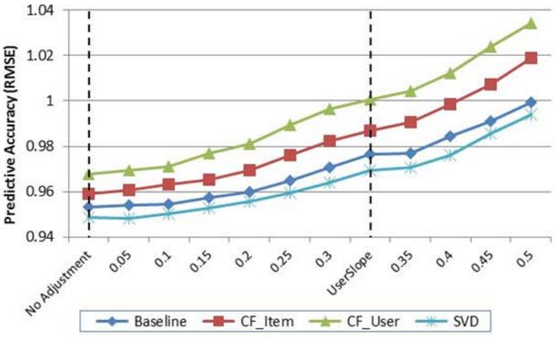


Fig 2. Predictive accuracy of de-biased ratings, with varying adjustment slopes.

Based on our experimental results, using users' actual submitted ratings (i.e., no adjustment) provided better accuracy performance than using de-biased ratings adjusted to any degree. As we increase the size of the global adjustment slope, the predictive accuracy performance estimated on test ratings decreases monotonically. Additionally, although the resulting accuracy of a user-specific adjustment is slightly better than that of the global slope of α = 0.35 suggested in prior research, the user-specific adjustment still did not yield better accuracy than no adjustment or small global adjustments. Overall, our experiment was unable to achieve any predictive accuracy improvements by de-biasing consumer ratings with either a global de-biasing rule based on a single slope parameter or the individual user-level rules based on user-specific slope parameters. We also conducted additional experiments with a variety of settings of post-hoc rating adjustment. For example, we introduced a tolerance threshold and only adjusted a submitted rating when it differs from the system's predicted rating by more than a certain amount (e.g., 0.5 stars). We also rounded de-biased ratings to various rating scales (e.g., to half stars, or to the first decimal place). We further experimented with adjusting only the positively biased ratings or only the negatively biased ratings to compare accuracy improvements. In addition, we empirically explored post-hoc rating de-biasing with a real-world movie rating dataset provided by Netflix [4].

However, based on our empirical explorations with these various post-hoc de-biasing methods, we have not been able to achieve any recommendation accuracy improvements by de-biasing consumer ratings with a global rule based on a single slope parameter (as demonstrated by Fig 2, we also explored other possible de-biasing slope values in addition to the empirically observed 0.35 value) or with a user-specific slope-based de-biasing rule. This indicates that, once the biased ratings are submitted, "reverse-engineering" is a difficult task. More specifically, while previous research was able to demonstrate that, *in aggregate*, there exist clear, measurable anchoring effects, it is highly likely that each *individual* anchoring effect (i.e., for a specific user/item rating) could be highly irregular – the biases could be user-dependent, item-dependent, context-dependent, and may have various types of other interaction effects. In fact, previous research provides some evidence to support such irregularity and situation-dependency. For example, prior studies observed symmetric (i.e., both positive and negative, equally pronounced) anchoring biases when they were aggregated across many items and asymmetric anchoring biases when they were tested on one specific item [1].

Therefore, an alternative approach to rating de-biasing would be to eliminate anchoring biases at rating-collection time through a carefully designed user interface. We discuss experiments with various interfaces in the next section.

4. APPROACH II: BIAS-AWARE INTERFACE DESIGN

The bias-aware interface design approach focuses on proactively preventing anchoring biases from occurring rather than trying to eliminate them after they have already occurred. We use a laboratory experiment to investigate various rating representation forms that may reduce anchoring effects at the rating collection stage. Besides the recommendation display, all other elements of the user interface were controlled to be equivalent across all experimental conditions. Our experiments explored *seven* different recommendation displays. Among them, four display designs were based on two main factors: (i) information representation (numeric vs. graphical ratings); and (ii) vagueness of recommendation (precise vs. vague rating values). Another two displays simulate popular star-rating representations used in many real-world recommender systems: stars-only and star along with a numeric rating. The seventh interface we explored was a binary design where only "thumbs up (down)" are displayed for high (low) predictions. Table 2 summarizes the seven rating representation options (i.e., Binary, Graphic-Precise, Graphic-Vague, Numeric-Precise, Numeric-Vague, Star-Numeric, and Star-Only).

4.1 Experiment Procedure

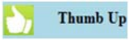
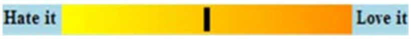
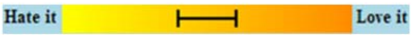
A database of 100 jokes was used for the study, with the order of the jokes randomized across participants. The jokes and the rating data for training the recommendation algorithm were taken from the Jester Online Joke Recommender System repository, a database of jokes and preference data maintained by the Univ. of California, Berkeley (http://eigentaste.berkeley.edu/dataset) [9]. The well-known item-based collaborative filtering technique was used to implement a recommender system that estimates users' preference ratings for the jokes [11]. The study was conducted at a behavioral research lab at a large North American university, and participants were recruited from the university's research participant pool. In total 287 people completed the study for a fixed participation fee.

Upon logging in, participants were randomly assigned to one of the seven treatment groups. Subjects in different treatment groups saw different displays of predicted rating. Examples of the display and number of participants in each treatment group are provided in Table 2.

The experimental procedure consisted of three tasks, all of which were performed using a web-based application on personal computers with dividers, providing privacy between participants.

Task 1. In the first task, each participant was asked to provide his/her preference ratings for 50 jokes randomly selected from the pool of 100 jokes. Ratings were provided using a scale from one to five stars with half-star increments, having the following verbal labels: * = "Hate it", ** = "Don't like it", *** = "Like it", **** = "Really like it", and ***** = "Love it". For each joke, we also asked participants to indicate whether they have heard the joke before. The objective of this joke-rating task was to capture joke preferences from the participants. Based on ratings provided in this task, predictions for the remaining unrated 50 jokes were computed.

Table 2. Example displays of system predicted ratings

Group	N	Example Display of Predicted Rating
Binary	40	 Thumb Up or Thumb Down
Graphic Precise	40	 Hate it Love it
Graphic-Vague	40	 Hate it Love it
Numeric-Precise	40	3.0 (out of 5)
Numeric-Vague	39	between 2.6 and 3.4 (out of 5)
Star-Numeric	45	★★★☆☆ 3.0 (out of 5)
Star-Only	43	★★★☆☆

Task 2. In the second task, from the remaining unrated 50 jokes, participants were presented with 25 jokes (using 5 recommendation conditions with 5 jokes each) along with a rating recommendation for each joke and 5 jokes without a recommendation (as a control condition). The recommendation conditions are summarized below:

- *High-Artificial*: randomly generated high recommendation between 3.5 and 4.5 stars (drawn from a uniform distribution)
- *Low-Artificial*: randomly generated low recommendation between 1.5 and 2.5 stars (drawn from a uniform distribution)
- *High-Perturbed*: algorithmic predictions were perturbed upward by 1 star
- *Low-Perturbed*: algorithmic predictions were perturbed downward by 1 star
- *Accurate*: actual algorithmic predictions (i.e., not perturbed)
- *Control*: no recommendation to act as a control

We first selected 5 jokes for the High-Perturbed condition and 5 jokes for the Low-Perturbed condition. These 10 jokes were chosen pseudo-randomly to assure that the manipulated ratings would fit into the 5-point rating scale. Among the remaining jokes we randomly selected 15 jokes and assigned them to three groups: 5 to Accurate, 5 to High-Artificial and 5 to Low-Artificial. 5 more jokes were added as a control with no predicted system rating provided. The 25 jokes with recommendations were randomly ordered and presented on five consecutive webpages (with 5 displayed on each page). The 5 control jokes were presented on the subsequent webpage. Participants were asked to provide their preference ratings for all these 30 jokes on the same 5-star rating scale.

Task 3. As the third task, participants completed a short survey that collected demographic and other individual information for use in the analyses.

4.2 Analysis and Results

The Perturbed vs. Artificial within-subjects manipulation described above represents two different approaches to the study of recommendation system bias. The Artificial recommendations provide a view of bias that controls for the value ranges shown, manipulating some to be high and some low, while not accounting for individual differences in preferences in providing the recommendations. The Perturbed recommendations control for such possible preference differences, allowing a view of recommendation error effects. We analyze the results from each of these approaches separately. First, we test different rating presentations with *artificially* (i.e. randomly) generated recommendations (i.e., not based on users' preferences).

4.2.1 Artificial Recommendations

Fig 3 presents a plot of the aggregate means of user-submitted ratings for each of the treatment groups when high and low artificial recommendations were provided. As can be seen in the figure, low artificial recommendations pull down user's preference ratings relative to the control, and the high artificial recommendations tend to increase user's preference ratings. As an initial analysis, for each rating display we performed pairwise *t*-tests to compare user submitted ratings after receiving high and low artificial recommendations. The *t*-test results are presented in Table 3.

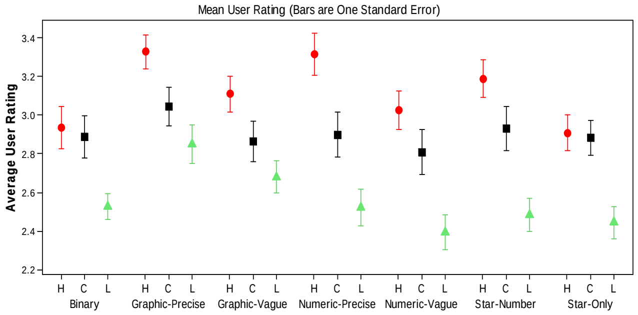


Fig 3. Mean and standard deviation of user submitted ratings after receiving high artificial (High: red dot), low artificial (Low: green triangle), or no recommendations (Control: black square).

Table 3. Pair-wise comparisons of mean user rating difference for each rating display option using t-tests.

Rating Display	High − Low	High − Control	Low − Control
Binary	0.408^{***}	0.045	-0.363^{***}
Graphic-Precise	0.478^{***}	0.283^{**}	-0.195^{*}
Graphic-Vague	0.428^{***}	0.245^{**}	-0.183^{*}
Numeric-Precise	0.793^{***}	0.415^{***}	-0.378^{***}
Numeric-Vague	0.628^{***}	0.215^{*}	-0.413^{***}
Star-Numeric	0.702^{***}	0.258^{***}	-0.444^{***}
Star-Only	0.463^{***}	0.026	-0.437^{***}

* p < 0.05, ** p < 0.01, ** p < 0.001

All comparisons between High and Low conditions are significant across the seven rating representations (one-tailed *p*-value < 0.001 for all High vs. Low tests), showing a clear, positive effect of randomly-generated recommendations on consumers' preference ratings. All effect sizes are large (Cohen's *d* values range between 0.71 and 1.23). The control condition demonstrated intermediate preference ratings, showing a statistically significant difference from the both High and Low conditions for the majority of the rating display options. This analysis demonstrates that the anchoring bias of artificial recommendations exists in *all* rating displays examined in our experiment. In other words, we found that none of the seven rating display options could completely remove the anchoring biases generated by recommendations.

We further compare the *anchoring bias size* of different rating display options. We computed rating differences between High and Low conditions and performed one-way ANOVA to test the overall group difference. Our results suggest significant difference in effect sizes among different rating representations ($F(6, 280) = 2.24$, $p < 0.05$). Since the overall effect was significant, we next performed regression analysis to explore the difference in anchoring bias between different rating display options, while controlling for participant-level factors.

In our regression analysis, we created a panel from the data. The repeated-measures design of the experiment, wherein each participant was exposed to both high and low artificial recommendations in a random fashion, allows us to model the aggregate relationship between shown ratings and user's submitted ratings while controlling for individual participant differences. The standard OLS model using robust standard errors, clustered by participant, and using participant-level controls represents our model for the analysis.

$$UserRating_{ij} = b_0 + b_1(Group_i) + b_2(High_{ij}) + b_3(Group_i \times High_{ij}) + b_4(ShownRatingNoise_{ij}) + b_5(PredictedRating_{ij}) + b_6(Controls) + u_i + \varepsilon_{ij}$$

In the regression equation shown above, $UserRating_{ij}$ is the submitted rating for participant *i* on joke *j*, $Group_i$ is the rating display option shown to participant *i*, $High_{ij}$ indicates whether the shown rating for participant *i* on joke *j* is a high or low artificial recommendations, $ShownRatingNoise_{ij}$ is a derived variable that captures the deviation between shown rating for participant *i* on joke *j* and the expected rating value in the corresponding condition. Specifically, it is computed by either subtracting 4.0 from the shown rating if it is in the high artificial condition or by subtracting 2.0 from the shown rating if it is in the low artificial condition. $PredictedRating_{ij}$ is the predicted recommendation star rating for participant *i* on joke *j*, and *Controls* is a vector of joke and consumer-related variables for participant *i*. The controls included in the model were the joke's funniness (average joke rating in the Jester dataset, continuous between 0 and 5), participant gender (binary), age (integer), 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 were useful (interval five point scale), and their self-reported numeracy levels reflecting participants' beliefs about their mathematical skills as a perceived cognitive ability using a scale of four items developed and validated by prior research [6] (continuous between 4 and 24). The latter information was collected in order to check for possible relationships between individual's subjective numeracy capabilities and individual's susceptibility to anchoring biases due to numeric vs. graphical rating displays. As 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}.

The Numeric-Precise rating display condition was chosen to be the baseline rating representation to compare with the other six options. We chose Numeric-Precise for two reasons. First it is a popular rating display used in many real-world recommender systems of large e-commerce websites such as Amazon, eBay and Netflix. Second, the Numeric-Precise rating display option was used by previous experiments in literature [1] and was found to lead to substantial anchoring biases in consumers' preference ratings. Therefore in our analysis we compare Numeric-Precise with other alternative rating display options to examine whether other rating representations can reduce the observed biases.

We ran three regression models with high artificial only, low artificial only, and both high and low artificial recommendations.

Note when only high or low recommendations were included for analysis, the model omitted the High variable and its related interaction terms. Table 4 presents the estimated coefficients and standard errors for the three regression models. All models utilized robust standard error estimates. The regression analysis controls for both participant and joke level factors as well as the participant's predicted preferences for the product being recommended.

Table 4. Regression analysis on artificial recommendations (baseline: Numeric-Precise; dependent variable: UserRating)

	Model 1	**Model 2**	**Model 3**
	High Only	Low Only	High&Low
Anchoring (High=1)			0.794***
ShownRatingNoise	0.350***	0.249**	0.289***
PredictedRating	0.319***	0.291***	0.289***
Group			
Binary	-0.372***	0.045	0.050
Graphic-Precise	-0.045	0.314**	0.301**
Graphic-Vague	-0.207*	0.176	0.165
Numeric-Vague	-0.238**	-0.073	-0.073
Star-Numeric	-0.149	-0.007	-0.015
Star-Only	-0.392***	-0.020	-0.036
Interactions			
Binary×Anchoring			-0.427***
Graphic-Precise×Anchoring			-0.331*
Graphic-Vague×Anchoring			-0.365*
Numeric-Vague×Anchoring			-0.169
Star-Numeric×Anchoring			-0.127
Star-Only×Anchoring			-0.345**
Controls			
jokeFunniness	0.618***	0.539***	0.587***
age	0.005	0.000	0.003
male	0.114*	0.009	0.063
native	-0.127*	-0.002	-0.067
PredictionAccurate	0.116***	0.005	0.062**
PredictionUseful	0.082***	-0.019	0.033
Numeracy	0.013	0.002	0.008
Constant	-2.219***	-0.592	-0.845***
R^2 within-subject	0.0514	0.0397	0.1485
R^2 between-subject	0.5735	0.3548	0.5561
R^2 overall	0.2648	0.1388	0.2450
χ^2	476.82***	155.74***	768.28***

* p < 0.05, ** p < 0.01, ** p < 0.001

Our analysis found randomly-generated recommendations displayed in Numeric-Precise format can substantially affect consumers' preference ratings, as indicated by significant coefficients for Anchoring and ShownRatingNoise in all three models. More importantly, we found significant negative interaction effects between multiple rating display options and anchoring (Model 3). The results clearly indicate that there are significant differences in anchoring biases between Numeric-Precise and other rating display options. Specifically, we observed that groups including Binary, Graphic-Precise, Graphic-Value, and Star-Only, when compared to Numeric-Precise, can generate much lower biases in consumers' preference ratings. All the corresponding interaction terms have negative coefficients with *p*-values smaller than 0.05. On the other hand, the interaction terms for Numeric-Vague and Star-Numeric were not significant, suggesting that these two display options lead to similar levels of anchoring biases as Numeric-Precise.

Overall, the Model 3 results suggest that, among all seven experimental rating display conditions, when randomly-assigned recommendations are presented in any non-numeric format (including Binary, Graphic-Precise, Graphic-Vague, Star-Only), they will generate much smaller anchoring biases compared to the same recommendations displayed in numeric formats such as Numeric-Precise, Numeric-Vague and Star-Numeric. In other words, the information representation of recommendations (e.g., numeric vs. non-numeric) largely determines the size of bias in consumers' preferences. Introducing vagueness to recommendations did not seem to reduce the anchoring bias when compared to the Numeric-Precise baseline (i.e., interaction between Numeric-Vague and anchoring is insignificant).

In a follow-up regression analysis (Table 5), we focused on four rating displays (i.e., Numeric-Precise, Numeric-Vague, Graphic-Precise, and Graphic-Vague) and similarly found the interaction between information presentation and anchoring (i.e., Numeric × Anchoring) was significant while the interaction between vagueness and anchoring (i.e., Precise × Anchoring) was not significant. This further confirms that the anchoring bias can be reduced by presenting recommendations in graphical forms rather than numeric forms. Anchoring bias, however, cannot be reduced by presenting the recommendations as vague rating ranges (as opposed to precise values).

Table 5. Regression analysis on artificial recommendations, for Numeric/Graphic and Precise/Vague rating displays (dependent variable: UserRating)

	Coefficient
Anchoring (High=1)	0.4027***
ShownRatingNoise	0.2024**
PredictedRating	0.2457***
Representation (Numeric=1)	-0.2667**
Vagueness (Precise=1)	0.1100
Numeric×Precise	0.0046
Numeric×Anchoring	0.2562**
Precise×Anchoring	0.1037
Controls	
jokeFunniness	0.7051***
age	0.0017
male	0.0788
native	-0.1013
PredictionAccurate	0.0687
PredictionUseful	0.0296
Numeracy	0.0146
Intercept	-1.0531**
R^2	0.2500
χ^2	420.37***

* p < 0.05, ** p < 0.01, ** p < 0.001

In addition, Model 1 focuses on high artificial recommendations (Table 4) and demonstrates significantly smaller anchoring biases for Binary, Graphic-Vague, Numeric-Vague and Star-Only displays, when compared to the Numeric-Precise display as the baseline. Model 2 focuses on low artificial recommendations and suggests that Graphic-Precise displays generated smaller biases compared to the baseline when recommendations were low. Therefore, another finding from Models 1 and 2 is that the "bias-reducing" effects of many rating display options can be highly asymmetric and depend on contextual factors such as the actual value of the recommendation.

Among the secondary factors, predicted consumer preferences, joke funniness, and perceived accuracy of recommendations all had consistently significant effects across all models. Therefore, controlling for these factors in the regression model was warranted.

4.2.2 Perturbed Recommendations

As an extension to a more realistic setting and as a robustness check, we next examine whether anchoring biases generated by *perturbations* in real recommendations from an actual recommender system can be eliminated by certain rating display options. Recall that participants received recommendations that were perturbed either upward (High-Perturbed) or downward (Low- Perturbed) by 1 star from the actual predicted ratings. As a control, each participant also received recommendations without perturbations (Accurate). Consumers' submitted ratings for the jokes were adjusted for the predicted ratings in order to obtain a response variable on a comparable scale across subjects. Thus, the main response variable is the rating drift, which we define as:

$$RatingDrift = UserRating - PredictedRating$$

Fig 4 is a plot of the aggregate means of rating drift for each treatment group when recommendations were perturbed to be higher or lower or received no perturbation. As can be seen, the negative perturbations (Low, green triangle) lead to negative rating drifts and positive perturbations (High, red dot) lead to positive drifts in user ratings, while the accurate recommendations with no perturbation (Accurate, black square) lead to drifts around zero. For each rating display, we performed pairwise *t*-tests to compare user-submitted ratings after receiving high and low artificial recommendations. The *t*-test results are presented in Table 6.

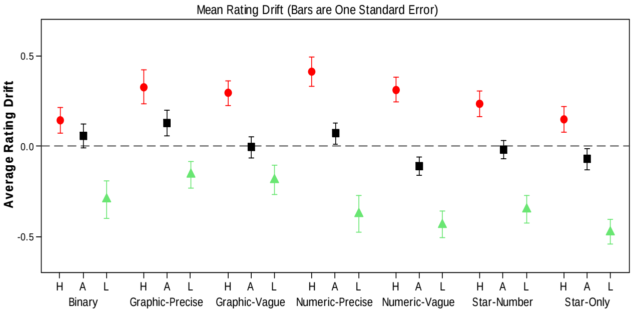


Fig 4. Mean and standard deviation of user rating drift after receiving high perturbed (High: red dot), low perturbed (Low: green triangle), and non-perturbed recommendations (Accurate: black square).

Table 6. Pairwise comparisons of mean rating drift difference for each rating display option using t-tests.

Rating Display	**High − Low**	**High − Accurate**	**Low − Accurate**
Binary	0.446^{***}	0.104	-0.318^{**}
Graphic-Precise	0.492^{***}	0.292^{**}	-0.187^{*}
Graphic-Vague	0.482^{***}	0.286^{**}	-0.196^{*}
Numeric-Precise	0.799^{***}	0.491^{***}	-0.297^{**}
Numeric-Vague	0.770^{***}	0.315^{**}	-0.420^{***}
Star-Numeric	0.599^{***}	0.196^{**}	-0.391^{***}
Star-Only	0.671^{***}	0.140^{*}	-0.474^{***}

* p < 0.05, ** p < 0.01, ** p < 0.001

All mean rating drift comparisons between High and Low perturbed conditions are significant for all rating display options (one-tailed *p*-value < 0.001 for all High vs. Low tests), showing a clear and positive anchoring bias of system recommendations on consumers' rating drift. Such anchoring biases exist in both High and Low perturbed conditions for the majority of the rating display options. The results clearly demonstrate that the anchoring effect of perturbed recommendations still exist in *all* rating display options investigated in our experiment. Hence, similar to the artificial groups, we found that none of the seven rating display options could completely remove the anchoring biases generated by perturbed real recommendations.

We next performed regression analysis to compare the size of anchoring bias across different rating display options, while controlling for participant-level factors. In our regression analysis, we created a panel from the data as each participant was exposed to both high and low perturbed recommendations in a random fashion. The standard OLS model using robust standard errors, clustered by participant, and participant-level controls represents our model for the analysis.

$$RatingDrift_{ij} = b_0 + b_1(Group_i) + b_2(High_{ij}) + b_3(Group_i \times High_{ij}) + b_4(PredictedRating_{ij}) + b_5(Controls) + u_i + \varepsilon_{ij}$$

In the above regression model, $RatingDrift_{ij}$ is the difference between submitted rating and predicted rating for participant i on joke j, $Group_i$ is the rating display option shown to participant i, $High_{ij}$ indicates whether the recommendation for participant i on joke j was perturbed upward or downward. *Controls* is the same vector of joke and consumer-related variables that was used in the previous regression analysis for artificial recommendations.

Table 7. Regression analysis on perturbed recommendations (baseline: Numeric-Precise; dependent variable: RatingDrift)

	Perturbed Recommendations High & Low
Anchoring (High = 1)	$0.777\ (0.119)^{***}$
PredictedRating	-0.128 (0.068)
Group	
Binary	0.081 (0.143)
Graphic-Precise	0.198 (0.126)
Graphic-Vague	0.159 (0.131)
Numeric-Vague	-0.087 (0.126)
Star-Numeric	0.023 (0.129)
Star-Only	-0.12 (0.126)
Interactions	
Binary×Anchoring	$-0.361\ (0.169)^{*}$
Graphic-Precise×Anchoring	-0.284 (0.168)
Graphic-Vague×Anchoring	$-0.302\ (0.152)^{*}$
Numeric-Vague×Anchoring	-0.042 (0.153)
Star-Numeric×Anchoring	-0.187 (0.157)
Star-Only×Anchoring	-0.139 (0.154)
Controls	
jokeFunniness	$0.236\ (0.095)^{**}$
age	0.002 (0.005)
male	0.016 (0.042)
native	-0.003 (0.052)
PredictionAccurate	0.032 (0.03)
PredictionUseful	0.011 (0.024)
Numeracy	0.011 (0.007)
Constant	-1.241 (0.405)
$\boldsymbol{R^2}$ within-subject	0.1493
$\boldsymbol{R^2}$ between-subject	0.0122
$\boldsymbol{R^2}$ overall	0.1214
χ^2	265.95^{***}

* p < 0.05, ** p < 0.01, ** p < 0.001

The regression model used ordinary least squares (OLS) estimation and a random effect to control for participant-level heterogeneity. The Numeric-Precise rating display condition was again chosen to be the baseline rating representation to compare with the other six options. Table 7 summarizes the regression analysis of perturbed recommendations.

Consistent with what we found in the artificial conditions, interaction terms between anchoring and some non-numeric displays including Binary and Graphic-Vague were significantly negative. Thus, when recommendations were displayed in Binary and Graphic-Vague formats, they generated much smaller rating drifts from consumer's actual preference, when compared to the baseline Numeric-Precise display.

Similar to Table 5, we also performed a 2×2 analysis on the two main dimensions: representation (numeric vs. graphic) and vagueness (precise vs. vague) of the displayed recommendations. Our results in Table 8 confirm that presenting recommendations in numeric format can lead to much larger ratings shifts in consumer's preference ratings than presenting the same recommendations in graphical format. The vagueness of recommendation value, however, does not have significant influence on size of anchoring bias.

Table 8. Regression analysis on perturbed recommendations, for Numeric/Graphic and Precise/Vague rating displays (dependent variable: RatingDrift)

	Coefficient
Anchoring (High=1)	0.4680***
PredictedRating	-0.1969*
Representation (Numeric=1)	-0.2558**
Vagueness (Precise=1)	0.0415
Numeric×Precise	0.0843
Numeric×Anchoring	0.2648*
Precise×Anchoring	0.0304
Controls	
jokeFunniness	0.4008
age	0.0097**
male	0.0975
native	-0.0381
PredictionAccurate	0.0779
PredictionUseful	-0.0378
Numeracy	0.0228
Intercept	-1.8631*
R^2	0.1497**
χ^2	420.37***

* p < 0.05, ** p < 0.01, ** p < 0.001

Overall, we observed that the real recommendations presented graphically can significantly lead to lower anchoring biases than real recommendations displayed in numeric forms (either as a precise number or as a numeric range). In addition, displaying real recommendations in binary format leads to much lower anchoring biases compared to recommendations in numeric forms (both numeric-precise and numeric-vague). Further, displaying real recommendations as a vague numeric range could not significantly reduce anchoring biases when compared to the benchmark approach of showing a precise value.

4.2.3 Discussion

Using several regression analyses and controlling for various participant-level factors, we found that none of the seven rating display options completely removed the anchoring biases generated by recommendations. However, we observed that some rating representations were more advantageous than others. For example, we find that graphical recommendations can lead to significantly lower anchoring biases than equivalent numeric forms (either as a precise number or a numeric range). In addition, displaying recommendations in binary format leads to lower anchoring biases compared to recommendations in numeric forms.

5. CONCLUSIONS

This paper focuses on the problem of "de-biasing" users' submitted preference ratings and proposes two possible approaches to remove anchoring biases from self-reported ratings.

The first proposed approach uses post-hoc adjustment rules to systematically sanitize user-submitted ratings that are known to be biased. We ran experiments under a variety of settings and explored both global adjustment rules and user-specific adjustment rules. Our investigation explicitly demonstrates the advantage of unbiased ratings over biased ratings on recommender systems' predictive performance. We also empirically show that post-hoc de-biasing of consumer preference ratings is a difficult task. Removing biases from submitted ratings using a global rule or user-specific rule is problematic, most likely due to the fact that the anchoring effects can manifest themselves very differently for different users and items. This further emphasizes the need to investigate more sophisticated post-hoc de-biasing techniques and, even more importantly, the need to proactively prevent anchoring biases in recommender systems during rating collection.

Therefore, the second proposed approach is a user-interface-based solution that tries to minimize anchoring biases at rating collection time. We provide several ideas for recommender systems interface design and demonstrate that using alternative representations can reduce the anchoring biases in consumer preference ratings. Using a laboratory experiment, we were not able to completely avoid anchoring biases with any of the variety of carefully designed user interfaces tested. However, we demonstrate that some interfaces are more advantageous for minimizing anchoring biases. For example, using graphic, binary, and star-only rating displays can help reduce anchoring biases when compared to using the popular numerical forms.

In future research, another possible de-biasing approach might be through consumer education, i.e., to make consumers more cognizant of the potential decision-making biases introduced through online recommendations. This constitutes an interesting direction for future explorations.

6. REFERENCES

[1] 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).

[2] Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2012. "Effects of Online Recommendations on Consumers' Willingness to Pay," *Conference on Information Systems and Technology*. Phoenix, AZ.

[3] Bell, R.M., and Koren, Y. 2007. "Improved Neighborhood-Based Collaborative Filtering," *KDDCup'07*, San Jose, CA, USA, 7-14.

[4] Bennet, J., and Lanning, S. 2007. "The Netflix Prize," *KDD Cup and Workshop*, www.netflixprize.com.

[5] Cosley, D., Lam, S., Albert, I., Konstan, J.A., and Riedl, J. 2003. "Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions," *CHI 2003 Conference*, Fort Lauderdale FL.

[6] Fagerlin, A., Zikmund-Fisher, B.J., Ubel, P.A., Jankovic, A., Derry, H.A., and Smith, D.M. 2007. "Measuring Numeracy without a Math Test: Development of the Subjective Numeracy Scale," *Medical Decision Making*, 27, 672-680.

[7] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. 1997. "Grouplens: Applying Collaborative Filtering to Usenet News," *Comm. the ACM*, 40, 77-87.

[8] Koren, Y., Bell, R., and Volinsky, C. 2009. "Matrix Factorization Techniques for Recommender Systems," *IEEE CS*, 42, 30-37.

[9] Lemire, D. 2005. "Scale and Translation Invariant Collaborative Filtering Systems," *Information Retrieval*, 8(1), 129-150.

[10] Sarwar, B., Karypis, G., Konstan, J.A., and Riedl, J. 2001. "Item-Based Collaborative Filtering Recommendation Algorithms," *Int'l WWW Conference*, Hong Kong, 285 - 295.

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