

Suggest or Sway? Effects of Online Recommendations on Consumers' Willingness to Pay

Gediminas Adomavicius¹, Jesse C. Bockstedt², Shawn P. Curley¹, Jingjing Zhang³

gedas@umn.edu, bockstedt@email.arizona.edu, curley@umn.edu, jjzhang@indiana.edu

¹ Information and Decision Sciences
Carlson School of Management
University of Minnesota
321 19th Avenue South
Minneapolis, MN 55455, USA

² Management Information Systems
Eller College of Management
The University of Arizona
McClellan Hall 430
1130 E. Helen Street
Tucson, Arizona 85721, USA

³ Operations and Decision Technologies
Kelley School of Business
Indiana University
1309 East Tenth Street
Business School, Room 670B
Bloomington, IN 47405, USA

Authors appear in alphabetical order.

Suggest or Sway?

Effects of Online Recommendations on Consumers' Willingness to Pay

ABSTRACT

Recommender systems have become an integral part of the online retail environment. Prior research has focused on advancing computational approaches to improving recommendation accuracy. Only recently have their behavioral implications become an object of study. In particular, the potential impact of recommender systems on *economic behavior* of individual consumers represents an important issue yet to be comprehensively explored. We present the results of two controlled experiments in the context of purchasing digital songs, which explore this issue. In the first study, we found strong evidence that randomly assigned song recommendations affected participants' willingness to pay, even when controlling for participants' preferences and demographics. In the second study, participants viewed actual system-generated recommendations that were intentionally perturbed (introducing recommendation error) and observed similar effects on willingness to pay. The results have significant implications for the design and application of recommender systems as well as for online retail practices.

Keywords: anchoring effects, behavioral economics, electronic commerce, experimental research, preferences, recommender systems

1. INTRODUCTION

Recommender systems are programs designed to assist consumers by presenting them with a subset of items that they are likely to find interesting, which is selected from a large set of possible items. The objective of a recommender system is to predict the preferences of a consumer (often expressed as numeric ratings) for items that they have not yet purchased, experienced or considered (depending on the specific application context). Recommender systems help to reduce search costs by protecting people from being overwhelmed by irrelevant and uninteresting information. Such systems have become commonplace in online purchasing environments, and personalized recommendations not only can add value to users' shopping experiences but can also be beneficial to sellers. For example, Amazon has reported that 35% percent of its product sales result from recommendations (Marshall 2006). Netflix, the Internet television and movie streaming/rental company, recently reported that about 75% of the content watched by its subscribers (including DVDs rented by mail and videos streamed online) was suggested by its recommendation system (Amatriain and Basilico 2012). Much research in information systems and computer science has focused on algorithmic design and improving recommender systems' performance (see Adomavicius and Tuzhilin 2005 for a review). However, very little research has explored the impact of recommender systems on consumer behavior from an economic or decision-making perspective. Considering how important recommender systems have become in purchase decision environments, there is a need to explore the influence of these systems on consumer economic behavior.

In this paper, we investigate this relationship by addressing the following general question: *Whether and to what extent do recommender system ratings that are displayed to users (indicating the system-estimated users' preference for items) influence users' economic behavior, i.e., their willingness to pay?* Drawing on theory from behavioral economics, judgment and decision-making, and marketing, we hypothesize that online recommendations¹ significantly pull a consumer's willingness to pay in the direction of the recommendation. We test our hypotheses with two controlled behavioral experiments

¹ In this paper, for ease of exposition, we use the term "recommendations" in a broad sense. Any rating that the consumer receives purportedly from a recommendation system, even if negative (e.g., 1 star on a five-star scale), is termed a recommendation of the system.

using the context of recommendations in the sale of popular songs. In the first study, we investigate whether randomly generated recommendations (i.e., not based on users' preferences) significantly impact consumers' willingness to pay for digital songs. In the second study, we extend these results and investigate whether the effects still exist for perturbations of actual recommendations generated by a real-time system that employs a popular, widely-used recommendation algorithm.

This work makes contributions along several lines. Unlike most of the anchoring research that we build upon and that uses tasks for which a correct answer exists now or in the near future, our recommendation setting deals with preference-based judgments, i.e., the judgment is a subjective preference and is not verifiable against an objective standard. Moreover, our focus is on the real economic impacts of the judgments being made. In addition, from the practical viewpoint, recommendation systems are prevalent and the influence of recommendations on consumers' judgments has several significant implications. In the next section, we detail our contributions in light of the previous research. Following that, we present two studies directed at the issue of interest: the economic impact of recommender systems' predictions. We conclude with a summary and discussion of potential issues arising from our observed results.

2. LITERATURE REVIEW AND HYPOTHESES

2.1. Anchoring Effects in Judgment and Decision Making

Prior research on judgment and decision making has shown that judgments can be constructed in real time and, as a consequence, they are often influenced by elements of the environment. One such influence arises from the use of an anchoring-and-adjustment heuristic (Tversky and Kahneman 1974; see review by Chapman and Johnson 2002), which provides the basis of the current study. When using this heuristic, the decision maker begins with an initial value (the anchor) and adjusts it as needed to arrive at the final judgment. A systematic bias has been observed with this process in that decision makers tend to arrive at a judgment that is skewed toward the initial anchor.

Past studies have largely been performed in artificial experimental settings not grounded in real-world

practices and using tasks for which a verifiable outcome is being judged, leading to a bias measured against an objective performance standard (e.g., Chapman and Johnson 2002). The recommendation setting of our research goes beyond past research along both of these dimensions.

The field of recommender systems provides a fertile arena for behavioral research because of the centrality of consumer behavior to the operation of these systems and the importance of these systems for online retailers. In this, the recommender systems fits within a still small but important body of research that “considers anchoring in all of its everyday variety and examines its various moderators in these diverse contexts” (Epley & Gilovich, 2010, p. 21). Other recent examples are works by Johnson, Schnytzer and Liu (2009)—who study anchoring in a real-world setting of horserace betting—and Ku, Galinsky and Murnighan (2006)—who investigate anchoring effects in auctions.

Investigating the role of recommender systems on judgment also contributes to the small body of work where the judgment is a subjective preference and is not verifiable against an objective standard (cf. Chapman and Johnson 2002). As noted by Ariely et al. (2003, p. 75), “the vast majority of anchoring experiments in the psychological literature have focused on how anchoring corrupts subjective judgment, not subjective valuation or preference.” Two of the few papers identified in the mainstream anchoring literature that have looked directly at anchoring effects in preference construction are those of Schkade and Johnson (1989) and Ariely et al. (2003). However, their work studied preferences in more abstract settings. Our research focuses on individuals working with a real recommender system (similar to those provided by online retailers with which consumers interact in their everyday life) and making judgments with real purchasing consequences.

In terms of anchoring effects on economic behavior, there has been little behavioral economics research. One example is the work by Ariely et al. (2006), who first asked students if they would pay the amount represented by the last two digits of their social security numbers for each of several items. Next, the researchers asked students to bid the maximum amount they would be willing to pay for each item and found that the initial anchor amount had significant influences on each student’s ultimate bids. Research has also found that charities can influence the amount donors give by manipulating how donation options are presented; people will give more if the suggested options are higher (Surowiecki 2004). In each case,

the provided anchors were in the same scale as consumers' responses, suggesting a possibility that anchoring arises from a scale compatibility effect, e.g., as observed with preference reversal phenomena (Tversky, Slovic & Kahneman 1990). In this present research, we eliminate the possible scale compatibility effect. As will be noted in the next section, we extend the results of prior research, which has shown that online recommendations introduce anchoring effects in consumers' preferences for items, to hypothesize that the anchoring effects on preference ratings also impact consumers' willingness to pay for those items.

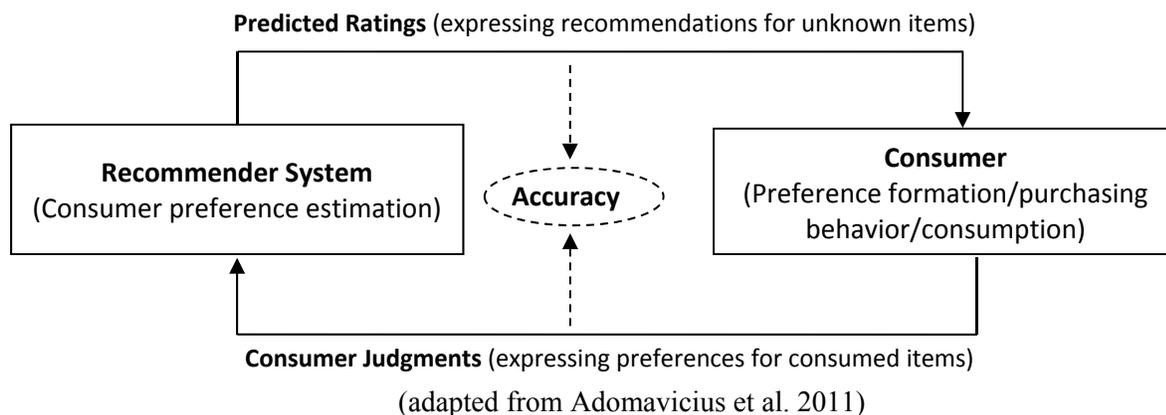
2.2. Anchoring and Recommender Systems

As illustrated in Figure 1, recommender systems often operate within a feedback loop. In most commercial recommender systems (e.g., Netflix, iTunes) consumers' preferences for items are modeled based on numeric ratings. Recommender systems take consumers' reported preferences (e.g., self-reported explicit numeric preference ratings for items they have experienced or implicit preferences imputed from related purchasing actions) as input and systematically predict consumers' ratings for not-yet-experienced items. These predicted ratings are often presented to consumers as recommendations to help them in their product purchase and item selection decisions. In turn, consumers' subsequently submitted ratings on newly consumed items are used as input for the system's estimation of future recommendations. These reported ratings are also commonly used to evaluate system performance by comparing how closely the system-predicted ratings match the users' reported ratings since the two sets of ratings are usually along the same scale (e.g., a 5-star rating scale). Studies corroborate that providing consumers with a predicted system rating introduces anchoring biases that significantly influence consumers' preference ratings (Cosley et al. 2003; Adomavicius et al. 2011). In the current studies, we investigate whether this feed-forward influence of the recommender system extends to actual economic behavior of the consumers.

Cosley et al. (2003) explored the effects of system-generated recommendations on user re-ratings of movies that they had seen in the past (i.e., not in the setting of the study). The re-ratings provided by users tended to be highly consistent with their original ratings when users were asked to re-rate a movie

with no preference prediction (i.e., recommendation) provided. 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, users tended to give higher or lower ratings, respectively (compared to a control group receiving accurate original ratings). This showed that anchoring could affect consumers’ ratings based on preference recall, for movies seen and rated in the past and now being re-evaluated. One explanation in non-preference settings for such anchoring effects upon preference recall is in terms of uncertainty of recall for queries like: How much did I enjoy that movie I saw last year? For example, Jacowitz & Kahneman (1995) argued that uncertainty about a quantity leads to a search from the anchor to the first plausible value in a distribution of uncertain values, leading to final estimates tilted toward the anchor.

Figure 1. Ratings as part of a feedback loop in consumer-recommender interactions.



More recently, Adomavicius et al. (2011) took a broader look at anchoring effects in the context of recommendations, including designs for which preference ratings were elicited at the time of consumption, thereby removing possible anchoring explanations operating at the point of preference recall. They also demonstrated a significant effect of system-generated recommendations on consumers’ preference ratings for items, now reported immediately after item consumption. In particular, using a within-subjects design, the researchers took the system-predicted ratings as a starting point and biased them (by perturbing them up or down) to varying degrees. The effect of perturbations upon ratings was

shown to be linear across positive and negative recommendations with a mean slope of 0.35; i.e., each unit shift in the shown system-generated rating resulted in a 0.35 unit change in self-reported consumer ratings. Thus, even without a delay between consumption and elicited preference (i.e., without any recall-related uncertainty), systematic and substantial anchoring effects were observed. The predicted rating, when perturbed to be higher or lower, affected the subsequently submitted consumer rating to move in the same direction. As identified by Cosley et al. (2003), these effects on consumer ratings are potentially important for a number of reasons, including: (1) biases can contaminate the inputs of the recommender system, reducing its effectiveness; (2) biases can artificially improve the resulting accuracy, providing a distorted view of the system's performance; or (3) biases might allow agents to manipulate the system so that it operates in their favor.

Beyond the effects upon preference and consumer ratings judgments, there is the question of effects upon the purchasing behavior of the consumer, as noted in Figure 1. The primary focus of the research presented here is to determine whether and to what extent the anchoring effects created by online recommendations upon preference ratings extend to impact consumers' economic behavior. Consumers receiving an artificially low or high recommendation are expected to express a willingness to pay more in line with the expectation in the corresponding direction. Thus, we expect the effects on preference ratings to extend to consumers' economic behavior, regardless of the accuracy of the recommendation.

H1: Participants exposed to randomly generated artificially high (low) recommendations for a product will exhibit a higher (lower) willingness to pay for that product.

A common issue for recommender systems is error (often measured by root mean square error, RMSE) in predicted ratings. For example, Netflix recently held a competition for a better recommendation algorithm with the goal of reducing prediction error, measured by RMSE, by 10% (Bennet and Lanning 2007). Adomavicius et al. (2011) and Cosley et al. (2003) demonstrated that anchoring biases in preference ratings result from inaccuracies (inadvertent or purposeful) in recommendations. We explore the economic impact when real recommendations (i.e., based on state-of-

the-art recommender system algorithms) are erroneous. We hypothesize that significant errors in real recommendations can affect consumers' behavior as captured by their willingness to pay for products.

H2: Participants exposed to a recommendation that contains error in an upward (downward) direction will exhibit a higher (lower) willingness to pay for the product.

The extension of the impact of system recommendations to willingness-to-pay judgments has a two-fold theoretical contribution. First is the extension to real economic behavior. Our participants make decisions among songs that will result in real purchases. Secondly, as noted in the previous section, previous work investigating anchoring effects of recommender systems (Cosley et al. 2003; Adomavicius et al. 2011) showed an impact of system recommendations provided on a 1-5 star scale upon consumer ratings on the same 5-star scale. A possibility under these conditions is that anchoring arises from a scale compatibility effect. The present study separates this correspondence. The system ratings are still provided on the 5-star rating scale; however subjects' economic behavior is measured along a completely different scale (US\$). Thus, to the extent that scale compatibility is essential to the anchoring effects with recommender systems observed in the earlier work, this would argue against Hypotheses 1 and 2. With this in mind, we test our hypotheses with two controlled behavioral studies, discussed next.

3. STUDY 1: RECOMMENDATIONS AND WILLINGNESS-TO-PAY

Study 1 was designed to test Hypothesis 1 and establish whether or not randomly generated recommendations could significantly impact a consumer's willingness to pay (WTP). Participants provided judgments for popular songs that were offered for purchase in digital format during the study.

3.1. Methods

3.1.1. Participants

The study was conducted at a behavioral research lab at a large public North American university, and participants were recruited from the university's research participant pool. Participants were paid a \$10 fee plus a \$5 endowment from which to purchase songs as described to them in the experimental

procedure (discussed below). Seven participants were dropped from Study 1 because of response issues: three subjects had zero variance in their WTP judgments and four subjects were deemed outside the scope of the general subject population. The final sample set consisted of 42 participants for analysis.

Demographic features of the sample are summarized in the first data column of Table 1. The participants were generally knowledgeable about music. Two-thirds of the participants indicated buying music at least once a month, with only seven stating that they never buy music. Nearly 3/4 of the participants said they owned more than 100 songs, with half (21/42) saying they own more than 1000 songs, and only one participant indicating that they own no songs.

Table 1. Participant Summary Statistics

	Study 1	Study 2
<i># of participants (n)</i>	42	55
<i>Average age (years)</i>	21.5 (1.95)	22.9 (2.44)
<i>Gender</i>	28 F, 14 M	31 F, 24 M
<i>Prior experience with recommender systems</i>	50% (21/42)	47.3% (26/55)
<i>Student level</i>	36 undergrad, 6 grad	27 undergrad, 25 grad, 3 other
<i>Buy new music at least once a month</i>	67% (28/42)	65% (36/55)
<i>Own more than 100 songs</i>	74% (31/42)	82% (45/55)

3.1.2. Stimuli

The stimuli were drawn from a database of 200 popular songs. The database consisted of songs in the bottom half of the year-end *Billboard* Hot 100 charts from 2006, 2007, 2008, and 2009.² These charts provide a mix of popular songs that we expected would be viewed favorably by the participants and that they would be willing to purchase. The bottom half of these charts were used because we were interested in WTP judgments only for songs that the participant did not already own (thereby removing current ownership as an influence on the judgments). It was expected that songs in the bottom half of the charts would be more likely to be both favorable and non-owned.

² Billboard Year-End Hot 100 Songs: <http://www.billboard.com/charts/year-end/2006/hot-100-songs>

3.1.3. Willingness to Pay (WTP) Judgments

To capture consumers' willingness to pay, we employed the incentive-compatible Becker, DeGroot, and Marschack (1964) method (BDM) commonly used in experimental economics. For each song presented during the third task of the study,³ participants were asked to declare a price they were willing to pay between zero and 99 cents. Participants were informed that five songs selected at random at the end of the study would be assigned random selling prices, based on a uniform distribution, between one and 99 cents. For each of these five songs, the participant was required to purchase the song using money from their \$5 endowment at the randomly assigned selling price if it was equal to or below their declared willingness to pay. Participants were presented with a detailed explanation of the BDM method so that they understood that the procedure incentivizes accurate reporting of their willingness to pay. Participants then took a quiz to check their knowledge of the procedure. 35/42 (83%) of participants answered the quiz questions perfectly on the first pass. Feedback was provided to correct mistaken answers before proceeding with the pricing task within the following procedure.

3.1.4. Procedure

The experimental procedure for Study 1 consisted of four main tasks, all of which were performed using a web-based application on personal computers with headphones and dividers, providing privacy between participants. The mean duration for participants to complete the session was just under 23 minutes, so fatigue was not judged to be an issue.

Task 1. In the first task, each participant was asked to provide his/her preference ratings for at least 50 songs randomly selected from the aforementioned pool of 200 songs. 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".⁴ For each song, the artist name(s), song title, duration, album name, and a 30-second sample were provided. Participants

³ The study had four tasks, which will be described in the next subsection.

⁴ This scale is similar to the scales used by common recommendation systems for rating entertainment items, such as the systems used by Netflix and Yahoo! Music.

could listen to any song at any time by clicking the link of the song sample, thereby reducing any recall uncertainty effects that might influence their responses. The objective of the song-rating task was to capture music preferences from the participants. This task provided a seeming basis for the system recommendations provided later in Study 1 (although the ratings were not used for this purpose) and as a real basis for the system recommendations in Study 2. These ratings also provide a basis for the post-hoc analysis of Study 1 to be discussed in the next section.

Task 2. In the second task, a different list of songs was presented (with the same information for each song as in the first task and with 20 songs on each screen), again randomly drawn for each participant from the same set of 200 songs but excluding the songs rated by the participant in Task 1. For each song, the participant was asked whether or not they owned the song until 40 non-owned songs were identified. Songs that were owned were excluded from the third task, in which willingness-to-pay judgments were obtained.

Task 3. In the third task of Study 1, participants first underwent training for the BDM pricing method as described in the previous subsection. They then completed a within-subjects experiment where the treatment was the star rating of the song recommendation and the dependent variable was willingness to pay for the songs. In the experiment, participants were presented with 40 songs that they did not own (from the second task), along with a star rating recommendation, artist name(s), song title, duration, album name, and a 30 second sample for each song. Participants were asked to specify their willingness to pay for each song on a scale from \$0.00 to \$0.99. The star rating recommendations were presented as personalized ratings for each participant. Ten of the 40 songs were presented with a randomly generated *low* recommendation between one and two stars (drawn from a uniform distribution; all recommendations were presented with a one decimal place precision, e.g., 1.3 stars), ten were presented with a randomly generated *high* recommendation between four and five stars, ten were presented with a randomly generated *mid-range* recommendation between 2.5 and 3.5 stars, and ten were presented with *no* recommendation to act as a control. The 30 songs presented with recommendations were randomly ordered and presented together on one webpage. The 10 control songs were presented on the following

webpage.

Task 4. As the fourth experimental task, participants completed a short survey that collected demographic and other individual information for use in the analyses. The participation fee and the endowment, minus fees paid for the required purchases, were distributed to participants in cash. MP3 versions of the songs purchased by participants were gifted to them through Amazon.com within 12 hours after the study was concluded.

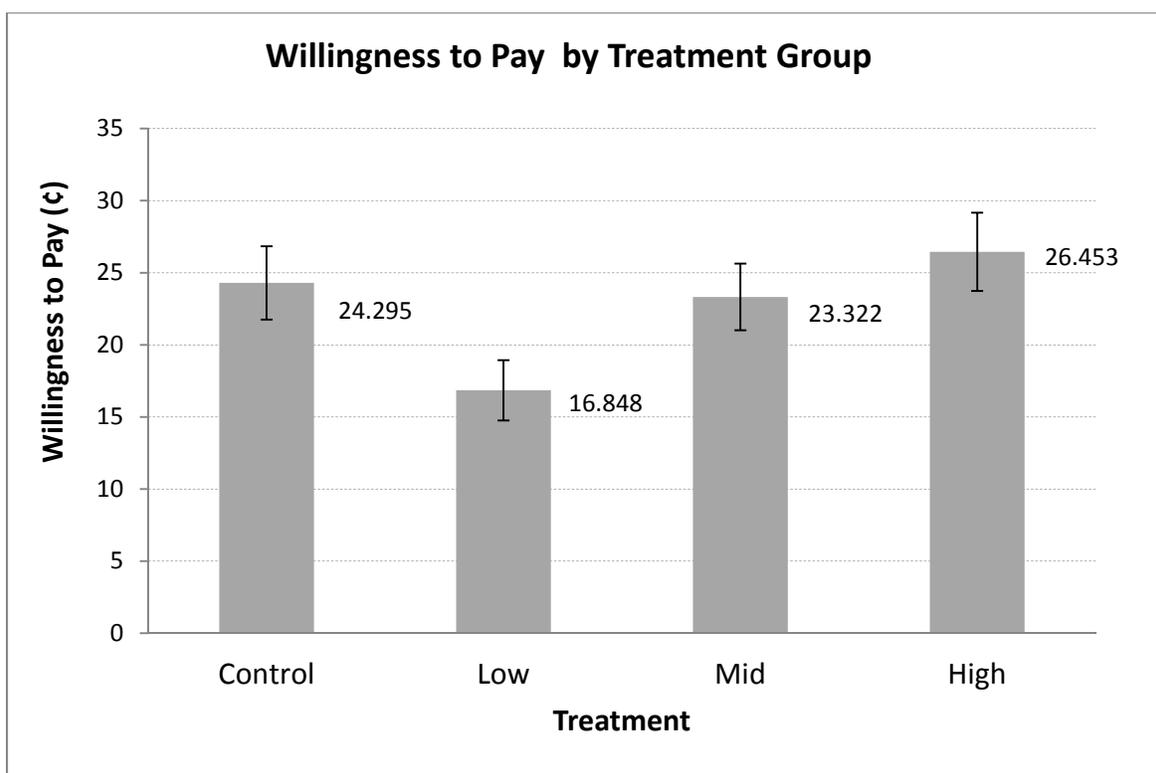
3.2. Results

Figure 2 shows a plot of the aggregate means of willingness to pay for each of the treatment groups. As an initial analysis, we performed a repeated measure ANOVA to test for differences across the three main treatment levels: High, Mid, and Low. The analysis demonstrated a statistically significant effect of the shown rating on willingness to pay ($F(2, 1196) = 42.27, p < .001$). Since the overall treatment effect was significant, we followed with pair-wise contrasts using t -tests across treatment levels and against the control group as shown in Table 2. All three main treatment conditions significantly differed, showing a clear, positive effect of the treatment on economic behavior. The control group demonstrated intermediate willingness to pay, showing a statistically significant difference from the Low treatment condition but not from the High treatment condition (one-tailed $p < .0001$ and $p = .13$, respectively) and no difference from the Medium condition (two-tailed $p = .58$).

We next performed a post hoc series of regression analyses to further explore the relationship between the shown star rating and willingness to pay, while controlling for participant-level factors. Note that, although there were three treatment groups, the actual ratings shown to the participants were randomly assigned star ratings from within the range for the corresponding treatment group (*low*: 1.0-2.0 stars, *mid*: 2.5-3.5 stars, *high*: 4.0-5.0 stars). Thus, in the regression analysis, the shown rating was a continuous variable ranging from 1.0-5.0 and was the main independent variable of interest. Control variables for several demographic and consumer-related factors were included, which were captured in the survey at the end of the study. Additionally, we controlled for the participants' preferences by calculating a predicted star rating recommendation for each song post hoc (on a 5-star scale with one

decimal precision), using the popular and widely-used item-based collaborative filtering algorithm (IBCF) (Sarwar et al. 2001).⁵ By including this predicted rating (which was not shown to any participant during Study 1) for each song-participant pair in the analysis, we are able to determine if the random recommendations had an impact on willingness to pay above and beyond the participant's predicted preferences.

Figure 2. Aggregate Treatment Means in Study 1



Note: Error bars represent 95% confidence intervals around mean of observations.

Table 2. Comparisons of Aggregate Treatment Group Means using t-Tests

	<i>Control</i>	<i>Low</i>	<i>Mid</i>
<i>Low</i> (1-2 Star)	4.436**		
<i>Mid</i> (2.5-3.5 Star)	0.555	4.075**	
<i>High</i> (4-5 Star)	1.138	5.501**	1.723*

* $p < 0.05$, ** $p < 0.01$, 2-tailed t-test for Control vs. Mid, all other contrasts were directional and tested with 1-tailed t-test.

⁵ Several recommendation algorithms were evaluated based on the Study 1 training data, and IBCF was selected by us in both studies because it had the highest predictive accuracy. More details on this aspect of the study are reported with Study 2.

The resulting baseline regression model is shown below, where WTP_{ij} is the reported willingness to pay for participant i on song j , $ShownRating_{ij}$ is the recommendation star-rating shown to participant i for song j , $PredictedRating_{ij}$ is the predicted recommendation star-rating for participant i on song j (i.e., the predicted preference), and $\mathbf{Controls}_i$ is a vector of demographic and consumer-related variables for participant i . The controls included in the model were: gender (binary); age (integer); school level (undergrad: yes/no binary); whether the participant has prior experience with recommendation systems (yes/no binary); preference ratings (interval five-point scale) for the music genres country, rock, hip hop, and pop; the number of songs owned (interval five-point scale); frequency of music purchases (interval five-point scale); whether they thought recommendations in the study were accurate (interval five-point scale); and whether they thought the recommendations were useful (interval five-point scale). Since the study utilized a repeated measures design with a balanced number of observations on each participant, the composite error term ($u_i + \varepsilon_{ij}$) includes an individual participant effect u_i in addition to the standard disturbance term ε_{ij} .

$$WTP_{ij} = b_0 + b_1(ShownRating_{ij}) + b_2(PredictedRating_{ij}) + b_3(\mathbf{Controls}_i) + u_i + \varepsilon_{ij} .$$

Three regression models were estimated and compared to account for the nature of the dependent variable. The baseline regression model (Model 1 in Table 3) used ordinary least squares (OLS) estimation. To control for participant-level heterogeneity, a random effect was used to model the individual participant effect. Random effects were chosen over fixed effects for three key reasons.⁶ First, we assume that the effects of the participants are randomly drawn from the overall population of potential participants. Second, the results of a Hausman test deemed the random effects model appropriate. Third, using random effects allows us to include participant-level controls in the analysis.

Prices often follow a log-normal distribution; therefore, we estimated a log-normal OLS regression with random participant-level effects in Model 2. Specifically, we used $\ln(WTP+1)$ as the dependent variable to account for the skewed distribution of the willingness-to-pay. For Model 3, a Tobit regression with participant-level random effects was estimated to account for potential censoring of the dependent

⁶ We also tested models with fixed participant-level effects in both studies and observed the same results in terms of significance, direction, and approximate magnitude.

variable (i.e., the participants' responses for willingness to pay were limited to a maximum value of 99 and a minimum value of 0). Tobit models are commonly used to model willingness-to-pay (e.g., Donaldson et al. 1998), and the assumption that zero-measured values of WTP reflect actual zero values and not a decision to enter the market holds in our controlled experimental setting. Table 3 presents the estimated coefficients and standard errors for the three models, and the results are highly consistent for the primary independent variables of interest, suggesting robustness of the results. All models utilized robust standard error estimates. Note that the control treatment observations were not included, since they had null values for the main independent variable *ShownRating*.

The results of our analysis for Study 1 provide strong support for H1 and demonstrate clearly that there is a significant effect of shown recommendations on consumers' economic behavior. Specifically, we observed that randomly generated recommendations with no dependence on actual user preferences can impact consumers' willingness to pay for a product. The regression analysis controls for participant-level factors and, most importantly, the participant's predicted preferences for the product being recommended. Looking first at the log-normal model (Model 2), we observed an increase (decrease) of 16% in willingness-to-pay for each 1-star increase (decrease) in the shown recommendation rating. The OLS model provides similar results: a 1-star increase (decrease) in the show recommendation results in a 3.5 cents US increase (decrease) in willingness to pay, in a sample with an average willingness to pay of approximately 20 cents US. The Tobit model provides similar results, although it should be noted that the coefficient for *ShownRating* (4.5110, $p \leq 0.001$) represents the marginal effect on the unobserved latent variable y^* , which represents the uncensored willingness to pay. Using the *margins* command in Stata 12, we computed the marginal effect for the conditional mean specification $E(\text{WillingnessToPay} \mid \mathbf{x}, 0 \leq \text{WillingnessToPay} \leq 99)$, where \mathbf{x} represents the collection of independent variables. This adjusted marginal effect takes into account censoring and was observed to be a 2.25 cents US ($p \leq 0.001$) increase (decrease) in willingness to pay for each 1-star increase (decrease) in the shown rating. Together, the regression results suggest that we can conservatively expect a positive effect of approximately 10-15% in willingness to pay for each 1-star increase in shown rating.

Table 3. Study 1 Regression Results, Dependent Variable: Willingness to Pay

	Model 1 OLS, RE	Model 2 LogNorm, RE	Model 3 Tobit, RE
ShownRating	3.5331*** (0.8007)	0.1618*** (0.0312)	4.5110*** (0.9823)
PredictedRating	6.2277*** (1.7171)	0.3953*** (0.1196)	9.4881*** (2.1751)
<i>Controls</i>			
male	-9.0832* (4.2883)	-0.8810** (0.2887)	-17.5101 (11.1667)
undergrad	-4.4357 (13.0865)	-0.3165 (0.740)	-3.1201 (19.4849)
age	-1.1651 (1.7336)	-0.0537 (0.1166)	-1.4913 (3.7016)
usedRecSys	-12.8029* (5.7752)	-1.1904*** (0.3728)	-20.6684* (10.2937)
country	2.8021 (1.5475)	0.2131* (0.1004)	4.1186 (3.4886)
rock	1.5246 (2.7020)	0.2906 (0.1643)	3.7363 (5.0877)
hiphop	-0.5736 (2.6187)	0.0945 (0.1491)	1.094 (4.8380)
pop	3.3729 (2.5347)	0.0901 (0.1819)	3.6954 (6.3411)
recomAccurate	-4.4175 (3.7294)	-0.2516 (0.2726)	-5.9156 (6.0102)
recomUseful	4.8637 (3.4587)	0.3185 (0.2536)	6.6289 (5.6126)
buyingFreq	-1.682 (2.6180)	-0.2270 (0.1466)	-3.325 (3.9904)
songsOwned	-2.85 (3.6847)	-0.2592 (0.2464)	-6.255 (7.1334)
constant	19.2 (51.7262)	1.7942 (3.3702)	7.2951 (99.5811)
R ²	0.21	0.27	0.21 (pseudo)
χ^2	87.4454***	139.80***	107.86***

Notes: Standard errors in parentheses, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, all models use robust standard error estimation. Model Summaries: Model 1 – ordinary least squares (OLS) estimation with random participant-level effects (RE); Model 2 – log-normal OLS (i.e., dependent variable = $\ln(\text{WTP}+1)$) with random participant-level effects; Model 3 – Tobit regression (upper limit 99, lower limit 0) with random participant-level effects. All models estimated using the Stata 12 software.

4. STUDY 2: ERRORS IN RECOMMENDATIONS

The goal of Study 2 was to extend the results of Study 1 by testing Hypothesis 2 and exploring the impact of errors of different magnitudes in true recommendations on consumers' willingness to pay. As discussed below, the design of this study is intended to test for similar effects as in Study 1, but in a more realistic setting with recommender system-generated, real-time recommendations.

4.1. Methods

4.1.1. Participants

Participants in Study 2 used the same facilities and were recruited from the same pool as in Study 1; however, there was no overlap in participants across the two studies. The same participation fee and endowment used in Study 1 was provided to participants in Study 2. Fourteen participants were dropped from Study 2 because of response issues: five subjects had zero variance in their WTP judgments and nine subjects were outside the scope of the general subject population. The final sample set consisted of 55 participants for analysis.

Demographic features of the sample are summarized in the second data column of Table 1. The participants are similar to those in Study 1 except for a lower percentage of undergraduates in the sample. Those in Study 2 were equally knowledgeable about music. Almost two-thirds of the participants indicated buying music at least once a month, with only four stating that they never buy music. More than 4/5 of the participants said they owned more than 100 songs, with nearly half (26/55) saying they own more than 1000 songs, and only one participant indicating that they own no songs.

4.1.2. Procedure

The stimuli database of 200 songs and the four tasks of the study were identical to Study 1. All participants completed the initial song-rating and song ownership tasks as in Study 1. Willingness-to-pay judgments were obtained using the BDM method with participant training and testing (with $36/55 = 65\%$ of participants answering the quiz questions perfectly on the first pass). The final survey, payouts, and song distribution were also conducted in the same manner as in Study 1. The mean duration for participants to complete the session was just under 27 minutes, similar in length as Study 1. The only difference between studies was in the design used for Task 3.

Study 2 again used a within-subjects design with willingness to pay as the dependent variable. However, in this case, the primary independent variable was the perturbation of a real recommender system's star-rating prediction for the participant. Unlike Study 1, the ratings that participants provided in Task 1 of Study 2 were used as inputs to a recommender system that determined predicted ratings for

the songs used in Tasks 2 and 3. In Study 2, real song recommendations were calculated based on the participants' preferences, which were then perturbed (i.e., excess error was introduced to each recommendation) to generate the shown recommendation ratings. Perturbations of -1.5 stars, -1 star, -0.5 stars, 0 stars, +0.5 stars, +1 star, and +1.5 stars were added to the actual recommendations, representing seven treatment levels. Each participant was presented with 40 songs for which WTP judgments were obtained, five from each of the seven treatment levels and five controls.

The 30 songs with perturbed ratings were determined pseudo-randomly to assure that the manipulated ratings would fit into the 5-point rating scale. First, 10 songs with predicted rating scores between 2.5 and 3.5 were selected randomly to receive perturbations of -1.5 and +1.5. From the remaining songs, 10 songs with predicted rating scores between 2.0 and 4.0 were selected randomly to receive perturbations of -1.0 and +1.0. Then, 10 songs with predicted rating scores between 1.5 and 4.5 were selected randomly to receive perturbations of -0.5 and +0.5. Finally, 5 songs were randomly selected and their predicted ratings were not perturbed; they were displayed exactly as predicted. These 35 songs were randomly intermixed. Following this, a final set of 5 songs were added as a control in random order (i.e., with no predicted system rating provided).

4.1.3. Recommender System

Based on the song rating data collected as Task 1 in Study 1, a recommender system was built for the purpose of making song recommendations in real time. We compared 6 popular recommendation techniques (Table 4) to find the best-performing technique for our dataset. The techniques included simple user- and item-based rating average methods, user- and item-based collaborative filtering (CF) approaches and their extensions (Bell and Koren 2007; Breese et al. 1998; Sarwar et al. 2001), as well as a model-based matrix factorization algorithm (Funk 2006; Koren et al. 2009) popularized by the recent Netflix prize competition (Bennett and Lanning 2007). Each technique was evaluated using the leave-one-out approach (Mitchell 1997) based on the standard root mean squared error (RMSE) and coverage metrics. Collaborative filtering (CF) algorithms performed best; and, consistent with the literature (Adomavicius and Tuzhilin 2005; Deshpande and Karypis 2004), item-based CF performed slightly better than the user-based CF approach. Based on these results, we used the item-based CF (IBCF) approach for

our recommender system (here and in Study 1).

Table 4. Comparison of Recommendation Techniques on Song Rating Dataset.

Methodology	Description	Predictive Accuracy (RMSE)	Predictive Coverage
Item Average	Predicts each user-item rating as an average rating of that item (or user, for user-based approach).	0.8936	1
User Average		0.8505	1
User-Item Average	Computes unknown ratings with baseline (i.e., “global effects”) estimates of corresponding users and items.	0.8190	1
Item-based CF	For each user-item rating to be predicted, finds the most similar items that have been rated by the user (or finds the most similar users who have rated the same item, for user-based approach) and computes the weighted sum of neighbors’ ratings as the predicted rating. Similarity is computed based on Pearson correlation coefficient (Konstan et al. 1997; Sarwar et al. 2001).	0.7790	1
User-based CF		0.7893	1
Matrix Factorization	Decomposes the rating matrix to two matrices so that every user and every item is associated with a user-factor vector and an item-factor vector. Prediction is done by taking the inner product of the user-factor and item-factor vectors.	0.8120	1

4.2. Results

As in Study 1, we start with a repeated measures ANOVA, which confirmed that there was a significant treatment effect of recommendation error (i.e., perturbation size) on willingness to pay ($F(5, 1649) = 8.64, p < .001$). Figure 3 presents the aggregate means by treatment condition. As can be seen in the figure, the negative perturbations pull down willingness to pay compared to the control, and the positive perturbations tend to increase willingness to pay.

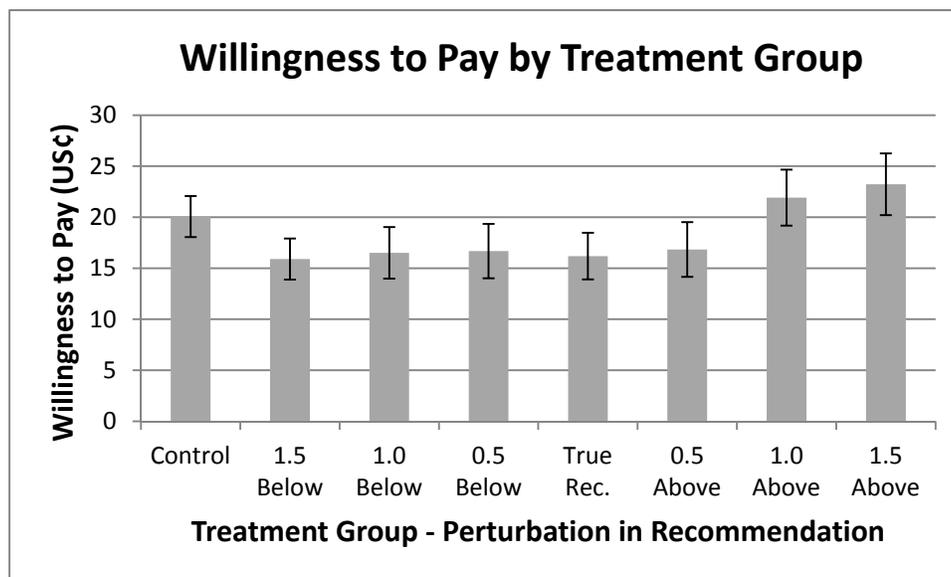
For our main analysis of Study 2, we used regression to examine the relationship between recommendation error and willingness to pay. We follow a similar approach as in Study 1 and analyze this relationship using three regression models. The distribution of willingness-to-pay data in Study 2 was similar to that of Study 1, so the same analysis strategy was adopted. The baseline model utilized OLS regression with random participant-level effects (the Hausman test for fixed versus random effects suggested that a random effects model was appropriate). We control for the participants’ preferences

using the predicted rating for each song in the study (i.e., the recommendation rating prior to perturbation), which was calculated using the IBCF algorithm. Furthermore, the same set of control variables used in Study 1 was included in our regression model for Study 2. The resulting regression model is presented below, where the main difference from the model used in Study 1 is the inclusion of $Perturbation_{ij}$ (i.e., the error introduced for the recommendation of song j to participant i) as the main independent variable.

$$WTP_{ij} = b_0 + b_1(Perturbation_{ij}) + b_2(PredictedRating_{ij}) + \mathbf{b}_3(\mathbf{Controls}_i) + u_i + \varepsilon_{ij}$$

As in Study 1, a log-normal OLS regression with random participant-level effects (Model 2) and a Tobit regression with random participant-level effects (Model 3) were each estimated and compared to account for the distribution of WTP. Robust standard errors were used in all models. The regression results are presented in Table 5.

Figure 3. Aggregate Treatment Group Means in Study 2



Note. Error bars represent 95% confidence intervals around mean of observations.

First, we note that although some of the control variables were significantly related to the dependent variable, WTP , no conclusions are drawn from these relationships. Study 1 (Table 3) also showed statistically significant results for several control variables. However, a comparison confirms that

there is no overlap between the two sets of significant variables across models or studies and, therefore, no consistent effect of the control variables.

Table 5. Study 2 Regression Results, Dependent Variable: Willingness to Pay

	Model 1 OLS, RE	Model 2 LogNorm, RE	Model 3 Tobit, RE
Perturbation	2.2447* (0.8800)	0.1211* (0.0493)	3.0223** (1.2402)
PredictedRating	8.7963*** (1.4142)	0.5675*** (0.082)	12.4401*** (1.9470)
<i>Controls</i>			
Male	-1.0053 (4.8348)	0.1966 (0.3173)	1.9951 (8.9549)
Undergrad	-4.8702 (4.5968)	-0.1566 (0.3176)	-6.4265 (8.7481)
Age	-0.3704 (0.8789)	-0.0157 (0.084)	-0.339 (1.4449)
usedRecSys	4.8136 (4.1486)	0.1961 (0.2781)	5.1331 (6.8373)
country	-2.2254 (1.5722)	-0.1088 (0.1209)	-2.6294 (3.0026)
rock	-2.339 (1.5456)	-0.2820** (0.0983)	-3.7576 (2.3417)
hiphop	-0.0264 (2.0705)	-0.0541 (0.1198)	-0.7016 (3.2784)
pop	-0.8571 (1.9623)	-0.0749 (0.1284)	-1.8523 (4.1371)
recomAccurate	1.2142 (2.2575)	0.0923 (0.1827)	0.352 (4.2785)
recomUseful	3.6726 (1.9526)	0.2029 (0.1419)	4.8032 (3.3032)
buyingFreq	3.0568* (1.4270)	0.1912 (0.1116)	2.4891 (2.6664)
songsOwned	-2.8066 (2.5310)	-0.6758 (0.1899)	-1.5455 (4.6081)
constant	2.8194 (23.4442)	-0.0741 (7.5478)	-4.5338 (40.3380)
R^2	0.16	0.17	0.21 (pseudo)
χ^2	71.0326***	121.36***	96.17***

Notes: Standard errors in parentheses, * $p \leq 0.05$, ** $p \leq 0.01$, *** $p \leq 0.001$, all models use robust standard error estimation. Model Summaries: Model 1 – ordinary least squares (OLS) estimation with random participant-level effects; Model 2 – log-normal OLS (i.e., dependent variable = $\ln(\text{WTP}+1)$) with random participant-level effects; Model 3 – Tobit regression (upper limit 99, lower limit 0) with random participant-level effects. All models estimated using the Stata 12 software.

As can be seen in Model 2 of Table 5, we observed a significant increase of approximately 12% in WTP for each 1-star positive increase in error of the shown recommendation, based on the log-normal regression model. The OLS model shows a 2.24 cents US ($p \leq 0.05$) increase in willingness to pay for

each 1-star positive increase in error of the shown recommendation. The Tobit model provided similar results; we observed a marginal effect of 3.02 cents US ($p \leq 0.05$) of the perturbation on the latent variable y^* , which is the unobserved and uncensored willingness to pay. The marginal effect for the conditional mean specification $E(\text{WillingnessToPay} \mid \mathbf{x}, 0 \leq \text{WillingnessToPay} \leq 99)$, where \mathbf{x} represents the collection of independent variables, is a 1.41 cents US ($p \leq 0.05$) increase in willingness to pay for each 1-star positive increase in error of the shown recommendation. The sample mean willingness to pay for Study 2 is 18.18 cents US; the regressions suggest that a significant effect corresponding to approximately 7-12% change in willingness to pay for each 1-star change in the perturbation can be expected.

The results of Study 2 provide strong support for H2 and extend the results of Study 1 in two important ways. First, Study 2 provides more realism to the analysis, since it utilizes real recommendations generated using an actual real-time system that applies a popular, commonly used recommendation algorithm. Second, rather than randomly assigning recommendations as in Study 1, in Study 2 the recommendations presented to participants were calculated based on their preferences and then perturbed to introduce realistic levels of system error. Considering the fact that all recommender systems have some level of error in their recommendations, Study 2 demonstrates the potential impact of these errors.

5. DISCUSSION

In two laboratory experiments we examined the impact of recommendations on consumers' economic behavior. The study integrates ideas from behavioral economics and recommender systems, both from practical and theoretical standpoints. The results provide strong evidence that predicted ratings provided by recommender systems can significantly influence consumer's economic behavior.

Study 1, through a randomized trial design, supported the hypothesis that online recommendations can affect willingness-to-pay judgments. Participants presented with random recommendations were influenced even when controlling for participant-level factors and preferences. Study 2 extended these results to demonstrate that the same effects exist for real recommendations that contain errors, using

recommendations that were calculated by applying the state-of-the-art recommendation algorithms used in practice.

From the theoretical perspective, the studies demonstrate evidence of anchoring-related effects in a realistic preference setting, whereas most prior anchoring research has been directed at judgments in response to general knowledge questions (as noted by Ariely et al. 2003; cf. Chapman and Johnson 2002). Moreover, our focus is on the real economic impacts of the judgments being made. Participant decisions in the experiment were made in a realistic setting with real economic consequences. Our studies thereby extend early work that investigated anchoring effects with recommender systems (e.g., Adomavicius et al. 2011; Cosley et al. 2003). Beyond the anchoring effects upon consumer ratings judgments, we substantiated the existence and the extent of comparable effects created by online recommendations upon consumers' economic behavior as measured by their willingness to pay.

In addition, the results of our studies, in combination with the prior work, indicate that the effect of recommender systems is not attributable to a scale compatibility effect, e.g., as observed with preference reversal phenomena (Tversky, Slovic & Kahneman 1990). In the present study, the system ratings are provided on the 5-star rating scale; however, subjects' economic behavior is measured along a completely different scale, 0-99 cents US.

There are also significant practical implications of the results presented. First, the observed results raise new issues regarding the design of recommender systems, which have been a popular tool for retailers and consumers. Can we reduce the biases that the recommendations introduce while maintaining the benefits that they provide? Applying general strategies for addressing biased judgments to the current setting suggests several possibilities.

One approach would be to mechanically adjust the recommendation algorithms used by consumers in judgments to correct the bias. Since recommender systems use a feedback loop based on consumer purchase decisions, it is open research question as to whether recommender systems should be calibrated to handle biased input to obtain a more accurate idea of performance. Another approach is to train consumers about the bias being exhibited. Consumers may need to become more cognizant of the potential decision-making biases introduced through online recommendations. Just as savvy consumers

understand the impacts of advertising, discounting, and pricing strategies, they may also need to consider the potential impact of recommendations on their purchasing decisions. On the flip side of this suggestion is the recognition that biases in decision making due to online recommendations can potentially be used to the advantage of economic agents. For example, retailers can potentially become more strategic in their use of recommender systems as a means of increasing profit and marketing to consumers. Yet another approach for de-biasing is to reconstruct the judgment interface to reduce the bias. This approach has been detailed by Thaler & Sunstein (2009) as the practice of decision or choice architecture, and is proving a promising area in general for the development of behavioral decision making and behavioral economics. Each of these three approaches is feasible in this context and need to be pursued in future research into recommender system use.

From the practical, economic perspective, we also note that we observed a 10-15% effect of a 1-star recommendation shift on willingness to pay in Study 1 and a 7-12% effect in Study 2. Many large online companies rely heavily on recommender systems in their retail practice (e.g., Amazon, iTunes, and Netflix). An effect of this size could have significant impact on revenues and profitability. As discussed earlier, Amazon has reported that 35% percent of its product sales result from recommendations (Marshall 2006). In 2011, Amazon net retail sales were approximately US\$46B;⁷ therefore, it would be reasonable to estimate that biases in consumer economic behavior due to recommendations at Amazon could potentially have an impact in the range of hundreds of millions of dollars. The true nature of these impacts on total welfare may not be so obvious. Recommendation errors that result in underestimating consumers' true preferences could potentially hurt online retailers, since lower recommendations would pull down consumers' willingness to pay for items. Alternatively, recommendation errors that result in inflated recommendation ratings could erode consumer surplus. In general, this issue of bias in recommendations is not trivial and can potentially have a net negative impact on online sales environments and the retail economy.

⁷ Global consolidated revenue not including services. For more details, see Amazon.com 2011 Annual report at <http://phx.corporate-ir.net/phoenix.zhtml?c=97664&p=irol-reportsannual>

Consequently, we strongly recommend pursuing further research in this area. As suggestions, we highlight three key areas as potential foci. First, field studies and analysis of secondary sales data from online retailers would help identify the true magnitude and impact of biases due to recommendations. Second, recommender system designs and implementations should be evaluated for potential impacts on consumer economic behavior from the standpoint of their decision architecture. There is a significant opportunity to rethink the way recommender systems calculate and present recommendations to users by taking into account the effects of system-induced biases. Finally, the evaluation of recommender systems may need to be reengineered to consider the potential for biases in the consumer preference input. If the inputs to the system are biased, simple comparisons of generated recommendations to reported preferences may no longer be the most effective means of measuring system performance.

REFERENCES

- Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2011. Recommender Systems, Consumer Preferences, and Anchoring Effects. *Proceedings of the RecSys 2011 Workshop on Human Decision Making in Recommender Systems*, Chicago IL, October 27, pp. 35-42.
- Adomavicius, G., and Tuzhilin, A. 2005. Towards the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17 (6) pp. 734-749.
- Amatriain, X. and Basilico, J. "Netflix Recommendations: Beyond the 5 stars," April 6, 2012. URL: <http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html>. Last accessed on June 9, 2013.
- Ariely, D., Lewenstein, G., and Prelec, D. "Coherent arbitrariness": Stable demand curves without stable preferences. *Quarterly Journal of Economics* (118) 2003, pp. 73-105.
- Ariely, D., Lowenstein, G., and Prelec, D. "Tom Sawyer and the construction of value," *Journal of Economic Behavior & Organization* 2006, pp 1-10.
- Becker G.M., DeGroot M.H., Marschak J. 1964. Measuring utility by a single-response sequential method. *Behavioral Science*, 9 (3) pp. 226–32.

- Bell, R.M., and Koren, Y. "Improved Neighborhood-based Collaborative Filtering," KDD Cup'07, San Jose, California, USA, 2007, pp. 7-14.
- Bennet, J., and Lanning, S. The Netflix Prize. *KDD Cup and Workshop*, 2007. [www.netflixprize.com].
- Breese, J.S., Heckerman, D., and Kadie, C. "Empirical analysis of predictive algorithms for collaborative filtering," Fourteenth Conference on Uncertainty in Artificial Intelligence, Madison, WI, 1998.
- Chapman, G., and Johnson, E. Incorporating the irrelevant: Anchors in judgments of belief and value. *Heuristics and Biases: The Psychology of Intuitive Judgment*, T. Gilovich, D. Griffin and D. Kahneman (eds.), Cambridge University Press, Cambridge, 2002, pp. 120-138.
- Cosley, D., Lam, S., Albert, I., Konstan, J.A., and Riedl, J. Is seeing believing? How recommender interfaces affect users' opinions. CHI 2003 Conference, Fort Lauderdale FL, 2003.
- Deshpande, M., and Karypis, G. "Item-Based Top-N Recommendation Algorithms," ACM Trans. Information Systems (22:1) 2004, pp. 143-177.
- Donaldson, C., Jones, A.M., Mapp, T.J., and Abel Olson, J. "Limited dependent variables in willingness to pay studies: applications in health care," *Applied Economics*, 30 (5) 667-677.
- Epley, N., and Gilovich, T. "Anchoring unbound," *J. of Consumer Psych.*, (20) 2010, pp. 20-24.
- Funk, S. "Netflix Update: Try This at Home," in: Netflix Update: Try This at Home, 2006. [http://sifter.org/~simon/journal/20061211.html].
- Johnson, J.E.V., Schnytzer, A., and Liu, S. "To what extent do investors in a financial market anchor their judgments excessively?" Evidence from the Hong Kong horserace betting market. *Journal of Behavioral Decision Making*, (22) 2009, pp. 410-434.
- Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. "GroupLens: Applying Collaborative Filtering to Usenet news," *Communications of the ACM* (40) 1997, pp 77-87.
- Koren, Y., Bell, R., and Volinsky, C. "Matrix Factorization Techniques For Recommender Systems," *IEEE Computer Society* (42) 2009, pp. 30-37.
- Ku, G., Galinsky, A.D., and Murnighan, J.K. "Starting low but ending high: A reversal of the anchoring effect in auctions," *J. of Personality and Social Psych.*, (90) 2006, pp. 975-986.
- Marshall, M. "Aggregate Knowledge raises \$5M from Kleiner, on a roll," December 10, 2006. URL: <http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/>. Last accessed on June 9, 2013.

- Mitchell, T. *Machine Learning* McGraw-Hill Science/Engineering/Math, 1997, p. 432.
- Sarwar B. Karypis, G., Konstan, J., Riedl, J. 2001. Item-based collaborative filtering algorithms. *10th Annual World Wide Web Conference (WWW10)*, May 1-5, Hong Kong.
- Schkade, D.A. and Johnson, E.J. Cognitive processes in preference reversals. *Organizational Behavior and Human Decision Processes*, (44) 1989, pp. 203-231.
- Surowiecki, J. *The wisdom of crowds : why the many are smarter than the few and how collective wisdom shapes business, economies, societies, and nations*, (1st ed.) Doubleday, New York, 2004, pp. xxi, 296 p.
- Thaler, R.H., and Sunstein, C.R. *Nudge: Improving decisions about health, wealth, and happiness* (rev. ed.). New York, Penguin Books, 2009.
- Tversky, A., and Kahneman, D. Judgment under uncertainty: Heuristics and biases. *Science*, (185) 1974, pp. 1124-1131.
- Tversky, A., Slovic P., and Kahneman, D. The causes of preference reversal. *American Economic Review*, (80) 1990, pp. 204-217.