Recommender Systems, Consumer Preferences, and Anchoring Effects

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

Jesse Bockstedt
George Mason University
Fairfax, VA
jbockste@gmu.edu

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

Jingjing Zhang
University of Minnesota
Minneapolis, MN
jingjing@umn.edu

ABSTRACT

Recommender systems are becoming a salient part of many e-commerce websites. Much research has focused on advancing recommendation technologies to improve the accuracy of predictions, while behavioral aspects of using recommender systems are often overlooked. In this study, we explore how consumer preferences at the time of consumption are impacted by predictions generated by recommender systems. We conducted three controlled laboratory experiments to explore the effects of system recommendations on preferences. Studies 1 and 2 investigated user preferences for television programs, which were surveyed immediately following program viewing. Study 3 broadened to an additional context—preferences for jokes. Results provide strong evidence viewers' preferences are malleable and can be significantly influenced by ratings provided by recommender systems. Additionally, the effects of pure number-based anchoring can be separated from the effects of the perceived reliability of a recommender system. Finally, the effect of anchoring is roughly continuous, operating over a range of perturbations of the system.

1. INTRODUCTION

Recommender systems have become important decision aids in the electronic marketplace and an integral part of the business models of many firms. Such systems provide suggestions to consumers of products in which they may be interested and allow firms to leverage the power of collaborative filtering and feature-based recommendations to better serve their customers and increase sales. In practice, recommendations significantly impact the decision-making process of many online consumers; for example, it has been reported that a recommender system could account for 10-30% of an online retailer's sales [25] and that roughly two-thirds of the movies rented on Netflix were ones that users may never have considered if they had not been recommended to users by the recommender system [10]. Research in the area of recommender systems has focused almost exclusively on the development and improvement of the algorithms that allow these systems to make accurate recommendations and predictions. Less well-studied are the behavioral aspects of using recommender systems in the electronic marketplace.

Many recommender systems ask consumers to rate an item that they have previously experienced or consumed. These ratings are then used as inputs by recommender systems, which employ various computational techniques (based on methodologies from statistics, data mining, or machine learning) to estimate consumer preferences for other items (i.e., items that have not yet been consumed by a particular individual). These estimated preferences are often presented to the consumers in the form of "system ratings," which indicate an expectation of how much the consumer will like the item based on the recommender system algorithm and, essentially, serve as recommendations. The subsequent consumer ratings serve as additional inputs to the system, completing a feedback loop that is central to a recommender system's use and value, as illustrated in Figure 1. The figure also illustrates how consumer ratings are commonly used to evaluate the recommender system's performance in terms of accuracy by comparing how closely the system-predicted ratings match the later submitted actual ratings by the users. In our studies, we focus on the *feed-forward* influence of the recommender system upon the consumer ratings that, in turn, serve as inputs to these same systems. We believe that providing consumers with a prior rating generated by the recommender system can introduce anchoring biases and significantly influence consumer preferences and, thus, their subsequent rating of an item. As noted by [7], biases in the ratings provided by users can lead to three potential problems: (i) biases can contaminate the inputs of the recommender system, reducing its effectiveness; (ii) biases can artificially improve the resulting accuracy, providing a distorted view of the system's performance; (iii) biases might allow agents to manipulate the system so that it operates in their favor.

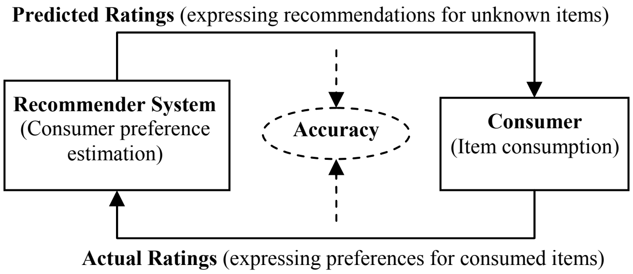
Predicted Ratings (expressing recommendations for unknown items)
Recommender System
(Consumer preference estimation)
Accuracy
Consumer
(Item consumption)
Actual Ratings (expressing preferences for consumed items)

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

For algorithm developers, the issue of biased ratings has been largely ignored. A common underlying assumption in the vast majority of recommender systems literature is that consumers have preferences for products and services that are developed independently of the recommendation system. However, researchers in behavioral decision making, behavioral economics, and applied psychology have found that people's preferences are often influenced by elements in the environment in which preferences are constructed [5,6,18,20,30]. This suggests that the common assumption that consumers have true, non-malleable preferences for items is questionable, which raises the following question: *Whether and to what extent is the performance of recommender systems reflective of the process by which preferences are elicited?* In this study, our main objective is to answer the above question and understand the influence of a recommender system's predicted ratings on consumers' preferences. In particular, we explore four issues related to the impact of recommender systems: (1) The *anchoring* issue—understanding any potential anchoring effect, particularly at the point of consumption, is the principal goal of this study: Are people's preference ratings for items they just consumed drawn toward predictions that are given to them? (2) The *timing* issue—

Decisions@RecSys'11, October 27, 2011, Chicago, IL, USA.

does it matter whether the system's prediction is presented before or after user's consumption of the item? This issue relates to one possible explanation for an anchoring effect. Showing the prediction prior to consumption could provide a prime that influences the user's consumption experience and his/her subsequent rating of the consumed item. If this explanation is operative, an anchoring effect would be expected to be lessened when the recommendation is provided *after* consumption. (3) The *system reliability* issue—does it matter whether the system is characterized as more or less reliable? Like the timing issue, this issue is directed at illuminating the nature of the anchoring effect, if obtained. If the system's reliability impacts anchoring, then this would provide evidence against the thesis that anchoring in recommender systems is a purely numeric effect of users applying numbers to their experience. (4) The *generalizability* issue—does the anchoring effect extend beyond a single context? We investigate two different contexts in the paper. Studies 1 and 2 observe ratings of TV shows in a between-subjects design. Study 3 addresses anchoring for ratings of jokes using a within-subjects-design. Consistency of our findings supports a more general phenomenon that affects preference ratings immediately following consumption, when recommendations are provided.

2. BACKGROUND AND HYPOTHESES

Behavioral research has indicated that judgments can be constructed upon request and, consequently, are often influenced by elements of the environment in which this construction occurs. One such influence arises from the use of an anchoring-and-adjustment heuristic [6,30], the focus of the current study. Using this heuristic, the decision maker begins with an initial value 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. Prior research on anchoring effects spans three decades and represents a very important aspect of decision making, behavioral economics, and marketing literatures. Epley and Gilovich [9] identified three waves of research on anchoring: (1) establishes anchoring and adjustment as leading to biases in judgment [5,9,21,29,30], (2) develops psychological explanations for anchoring effects [5,9,13,21,23], and (3) unbinds anchoring from its typical experimental setting and "considers anchoring in all of its everyday variety and examines its various moderators in these diverse contexts" ([9], p.21) [14,17]. Our study is primarily located within the latter wave while informing the second wave—testing explanations—as well; specifically, our paper provides a contribution both (a) to the study of anchoring in a preference situation at the time of consumption and (b) to the context of recommender systems.

Regarding the former of these contextual features, the effect of anchoring on preference construction is an important open issue. Past studies have largely been performed using tasks for which a verifiable outcome is being judged, leading to a bias measured against an objective performance standard (also see review by [6]. In the recommendation setting, the judgment is a *subjective* preference and is not verifiable against an objective standard. The application of previous studies to the preference context is not a straightforward generalization.

Regarding our studies' second contextual feature, very little research has explored how the cues provided by recommender systems influence online consumer behavior. The work that comes closest to ours is [7], which explored the effects of system-generated recommendations on user re-ratings of movies. It found that users showed high test-retest consistency when being asked to re-rate a movie with no prediction provided. However, when users were asked to re-rate a movie while being shown a "predicted" rating that was altered upward/downward from their original rating for the movie by a single fixed amount (1 rating point), they tended to give higher/lower ratings, respectively.

Although [7] did involve recommender systems and preferences, our study differs from theirs in important ways. First, we address a fuller range of possible perturbations of the predicted ratings. This allows us to more fully explore the anchoring issue as to whether any effect is obtained in a discrete fashion or more continuously over the range of possible perturbations. More fundamentally, the focus of [7] was on the effects of anchors on a *recall* task, i.e., users had already "consumed" (or experienced) the movies they were asked to re-rate in the study, had done so prior to entering the study, and were asked to remember how well they liked these movies from their past experiences. Thus, anchoring effects were moderated by potential recall-related phenomenon, and preferences were being remembered instead of constructed. In contrast, our work focuses on anchoring effects that occur in the construction of preferences at the time of actual consumption. In our study, no recall is involved in the task impacted by anchors, participants consume the good for the first time in our controlled environment, and we measure the immediate effects of anchoring.

Still, [7] provide a useful model for the design of our studies, with two motivations in mind. First, their design provides an excellent methodology for exploring the effects of recommender systems on preferences. Second, we build upon their findings to determine if anchoring effects of recommender systems extend beyond recall-related tasks and impact actual preference construction at the time of consumption. Grounded in the explanations for anchoring, as discussed above, our research goes beyond their findings to see if recommender system anchoring effects are strong enough to manipulate a consumer's perceptions of a consumption experience as it is happening.

Since anchoring has been observed in other settings, though different than the current preference setting, we begin with the conjecture that the rating provided by a recommender system serves as an anchor. Insufficient adjustment away from the anchor is expected to lead to a subsequent consumer preference rating that is shifted toward the system's predicted rating. This is captured in the following primary hypothesis of the studies:

> Anchoring Hypothesis: Users receiving a recommendation biased to be higher will provide higher ratings than users receiving a recommendation biased to be lower.

One mechanism that may underlie an anchoring effect with recommendations is that of *priming*, whereby the anchor can serve as a prime or prompt that activates information similar to the anchor, particularly when uncertainty is present [6]. If this dynamic operates in the current setting, then receiving the recommendation prior to consumption, when uncertainty is higher and priming can more easily operate, should lead to greater anchoring effects than receiving the recommendation after consumption. Manipulating the timing of the recommendation provides evidence for tying any effects to priming as an underlying mechanism.

> Timing Hypothesis: Users receiving a recommendation prior to consumption will provide ratings that are closer to the recommendation (i.e., will be more affected by the anchor) than users receiving a recommendation after viewing.

Another explanation proposed for the anchoring effect is a content-based explanation, in which the user perceives the anchor as providing evidence as to a correct answer in situations where an objective standard exists. When applied to the use of recommender systems and preferences, the explanation might surface as an issue of the consumer's trust in the system. Prior study found that increasing cognitive trust and emotional trust improved consumer's intentions to accept the recommendations [15]. Research also has highlighted the potential role of human-computer interaction and system interface design in achieving high consumer trust and acceptance of recommendations [7,19,22,28]. However, the focus of these studies differs from that underlying our research questions. In particular, the aforementioned prior studies focused on interface design (including presentation of items, explanation facilities, and rating scale definitions) rather than the anchoring effect of recommendations on the construction of consumer preferences. Our work was motivated in part by these studies to specifically highlight the role of *anchoring* on users' preference ratings.

In their initial studies, Tversky and Kahneman [30] used anchors that were, explicitly to the subjects, determined by spinning a wheel of fortune. They still observed an effect of the magnitude of the value from this random spin upon the judgments made (for various almanac-type quantities, e.g., the number of African countries in the United Nations). [27] also demonstrated anchoring effects even with extreme values (e.g., anchors of 1215 or 1992 in estimating the year that Einstein first visited the United States). These studies suggest that the anchoring effect may be purely a numerical priming phenomenon, and that the quality of the anchor may be less important.

In contrast, [20] found that the anchoring effect was mediated by the plausibility of the anchor. The research cited earlier connecting cognitive trust in recommendation agents to users' intentions to adopt them [15] also suggests a connection between reliability and use. To the extent that the phenomenon is purely numerically driven, weakening of the recommendation should have little or no effect. To the extent that issues of trust and quality are of concern, a weakening of the anchoring should be observed with a weakening of the perceived quality of the recommending system.

> Perceived System Reliability Hypothesis: Users receiving a recommendation from a system that is perceived as more reliable will provide ratings closer to the recommendation (i.e., will be more affected by the anchor) than users receiving a recommendation from a less reliable system.

To explore our hypotheses, we conducted three controlled laboratory experiments, in which system predictions presented to participants are biased upward and downward so our hypotheses can be tested in realistic settings. The first study explores our hypotheses by presenting participants with randomly assigned artificial system recommendations. The second study extends the first and uses a live, real-time recommender system to produce predicted recommendations for our participants, which are then biased upward or downward. The final study generalizes to preferences among jokes, studied using a within-subjects design and varying levels of rating bias. The next three sections provide details about our experiments and findings.

3. STUDY 1: IMPACT OF ARTIFICIAL RECOMMENDATIONS

The goals of Study 1 were fivefold: (1) to perform a test of the primary conjecture of anchoring effects (i.e., Anchoring Hypothesis) using artificial anchors; (2) to perform the exploratory analyses of whether participants behave differently with high vs. low anchors; (3) to test the Timing Hypothesis for anchoring effects with system recommendations (i.e., concerning differential effects of receiving the recommendation either before or after consuming the item to be subsequently rated) ; (4) to test the Perceived System Reliability Hypothesis for anchoring effects with system recommendations (i.e., concerning the relationship between the perceived reliability of the recommender system and anchoring effects of its recommendations); and (5) to build a database of user preferences for television shows, which would be used in computing personalized recommendations for Study 2.

3.1. Methods

216 people completed the study. Ten respondents indicated having seen some portion of the show that was used in the study (all subjects saw the same TV show episode in Study 1). Excluding these, to obtain a more homogeneous sample of subjects all seeing the show for the first time, left 206 subjects for analysis. Participants were solicited from a paid subject pool and paid a fixed fee at the end of the study.

In Study 1 subjects received *artificial* anchors, i.e., system ratings were not produced by a recommender system. All subjects were shown the same TV show episode during the study and were asked to provide their rating of the show after viewing. Participants were randomly assigned to one of seven experimental groups. Before providing their rating, those in the treatment groups received an artificial system rating for the TV show used in this study. Three factors were manipulated in the rating provision. First, the system rating was set to have either a *low* (1.5, on a scale of 1 through 5) or *high* value (4.5). Since [29] found an asymmetry of the anchoring effect such that high anchors produced a larger effect than did low anchors in their study of job performance ratings, we used anchors at both ends of the scale.

The second factor in Study 1 was the timing of the recommendation. The artificial system rating was given either *before* or *after* the show was watched (but always before the viewer was asked to rate the show). This factor provides a test of the Timing Hypothesis. Together, the first two factors form a 2 x 2 (High/Low anchor x Before/After viewing) between-subjects design (the top four cells of the design in Table 1).

Intersecting with this design is the use of a third factor: the perceived reliability of the system (*strong* or *weak*) making the recommendation. In the Strong conditions for this factor, subjects were told (wording is for the Before viewing/Low anchor condition): "Our recommender system thinks that you would rate the show you are about to see as 1.5 out of 5." Participants in the corresponding Weak conditions for the perceived reliability factor saw: "We are testing a recommender system that is in its early stages of development. Tentatively, this system thinks that you would rate the show you are about to see as 1.5 out of 5." This factor provides a test of the Perceived System Reliability Hypothesis. At issue is whether any effect of anchoring upon a recommendation is merely a numerical phenomenon or is tied to the perceived reliability and quality of the recommendation.

Since there was no basis for hypothesizing an interaction between timing of the recommendation and strength of the system, the complete factorial design of the three factors was not employed. For parsimony of design, the third factor was manipulated only within the Before conditions, for which the system recommendation preceded the viewing of the TV show. Thus, within the Before conditions of the Timing factor, the factors of Anchoring (High/Low) and Reliability of the anchor (Strong/Weak) form a 2x2 between-subjects design (the bottom four cells of the design in Table 1).

In addition to the six treatment groups, a control condition, in which no system recommendation was provided, was also included. The resulting seven experimental groups, and the sample sizes for each group, are shown in Table 1.

Subjects participated in the study using a web-based interface in a behavioral lab, which provided privacy for individuals participating together. Following a welcome screen, subjects were shown a list of 105 popular, recent TV shows. TV shows were listed alphabetically within five genre categories: Comedy, Drama, Mystery/Suspense, Reality, and Sci Fi/Fantasy. For each show they indicated if they had ever seen the show (multiple episodes, one episode, just a part of an episode, or never), and then rated their familiarity with the show on a 7-point Likert scale ranging from "Not at all familiar" to "Very familiar." Based on these responses, the next screen first listed all those shows that the subject indicated having seen and, below that, shows they had not seen but for which there was some familiarity (rating of 2 or above). Subjects rated each of these shows using a 5-star scale that used verbal labels parallel to those in use by Netflix.com. Half-star ratings were also allowed, so that subjects had a 9-point scale for expressing preference. In addition, for each show, an option of "Not able to rate" was provided. Note that these ratings were not used to produce the artificial system recommendations in Study 1; instead, they were collected to create a database for the recommender system used in Study 2 (to be described later).

Table 1 Experimental Design and Sample Sizes in Study 1.

Control: 29			
Reliability condition	Timing condition	Low (anchor)	High (anchor)
Strong (reliability)	After (timing)	29	28
Strong (reliability)	Before (timing)	29	31
Weak (reliability)	Before (timing)	29	31

Following the rating task, subjects watched a TV episode. All subjects saw the same episode of a situation comedy. A less well-known TV show was chosen to maximize the likelihood that the majority of subjects were not familiar with it. The episode was streamed from Hulu.com and was 23 minutes 36 seconds in duration. The display screen containing the episode player had a visible time counter moving down from 20 minutes, forcing the respondents to watch the video for at least this time before the button to proceed to the next screen was enabled.

Either immediately preceding (in the Before conditions) or immediately following (in the After conditions) the viewing display, subjects saw a screen providing the system recommendation with the wording appropriate to their condition (Strong/Weak, Low/High anchor). This screen was omitted in the Control condition. Following, subjects rated the episode just viewed. The same 5-star (9-point) rating scale used earlier was provided for the preference rating, except that the "Not able to rate" option was omitted. Finally, subjects completed a short survey that included questions on demographic information and TV viewing patterns.

3.2. Results

All statistical analyses were performed using SPSS 17.0. Table 2 shows the mean ratings for the viewed episode for the seven experimental groups. Our preliminary analyses included data collected by survey, including both demographic data (e.g., gender, age, occupation) and questionnaire responses (e.g., hours watching TV per week, general attitude towards recommender systems), as covariates and random factors. However, none of these variables or their interaction terms turned out to be significant, and hence we focus on the three fixed factors.

We begin with analysis of the 2x2 between-subjects design involving the factors of direction of anchor (High/Low) and its timing (Before/After viewing). As is apparent from Table 2 (rows marked as Design 1), and applying a general linear model, there is no effect of Timing (*F*(1,113) = 0.021, *p* = .885). The interaction of Timing and High/Low anchor was also not significant (*F*(1, 113) = 0.228, *p* = .634). There is a significant observed anchoring effect of the provided artificial recommendation (*F*(1, 113) = 14.30, *p* = .0003). The difference between the High and Low conditions was in the expected direction, showing a substantial effect between groups (one-tailed *t*(58) = 2.788, *p* = .0035, assuming equal variances). Using Cohen's (1988) *d,* which is an effect size measure used to indicate the standardized difference between two means (as computed by dividing the difference between two means by a standard deviation for the data), the effect size is 0.71, in the medium-to-large range.

Table 2. Mean (SD) Ratings of the Viewed TV Show by Experimental Condition in Study 1.

Design 1	Design 2	Group (timing-anchor-reliability)	N	Mean (SD)
*	*	Before-High-Strong	31	3.48 (1.04)
*		After-High-Strong	28	3.43b (0.81)
		Control	29	3.22 (0.98)
	*	Before-High-Weak	31	3.08 (1.07)
	*	Before-Low-Weak	29	2.83 (0.75)
*		After-Low-Strong	29	2.88 (0.79)
*	*	Before-Low-Strong	29	2.78 (0.92)

Using only data within the Before conditions, we continue by analyzing the second 2 x 2 between-subjects design in the study (Table 2, rows marked as Design 2), involving the factors of direction of anchor (High/Low) and perceived system reliability (Strong/Weak). The anticipated effect of weakening the recommender system is opposite for the two recommendation directions. A High-Weak recommendation is expected to be less pulled in the positive direction compared to a High-Strong recommendation; and, a Low-Weak recommendation is expected to be less pulled in the negative direction as compared to Low-Strong. So, we explore these conjectures by turning to the direct tests of the contrasts of interest. There is no significant difference between the High and Low conditions with Weak recommendations (*t*(58) = 1.053, *p* = .15), unlike with Strong recommendations (as noted above, p = .0035). Also, the overall effect was reduced for the Weak setting, compared to the Strong recommendation setting, and was measured as a Cohen's *d* = 0.16, less than even the small effect size range. Thus, the subjects were sensitive to the perceived reliability of the recommender system. Weak recommendations did not operate as a significant anchor when the perceived reliability of the system was lowered.

Finally, we check for asymmetry of the anchoring effect using the control group in comparison to the Before-High and Before-Low groups. (Similar results were obtained using the After-High and After-Low conditions as comparison, or using the combined High and Low groups.) In other words, we already showed that the High and Low groups were significantly different from each other, but we also want to determine if each group differs from the Control (i.e., when no recommendation was provided to the users) in the same manner. When an artificial High recommendation was provided (4.5), ratings were greater than those of the Control group, but not significantly so ($t(58) = 0.997$, $p = .162$). But when an artificial Low recommendation was provided (1.5), ratings were significantly lower than those of the Control group ($t(56) = 1.796$, p = .039). There was an asymmetry of the effect; however, the direction was opposite to that found by Thorsteinson et al. (2008). To study the effect further, Study 2 was designed to provide further evidence. So, we will return to the discussion of the effect later in the paper.

In summary, analyses indicate a moderate-to-strong effect, supporting the Anchoring Hypothesis. When the recommender system was presented as less reliable, being described as in test phase and providing only tentative recommendations, the effect size was reduced to a minimal or no effect, in support of the Perceived System Reliability Hypothesis. Finally, the Timing Hypothesis was not supported – the magnitude of the anchoring effect was not different whether the system recommendation was received before or after the viewing experience. This suggests that the effect is not attributable to a priming of one's attitude prior to viewing. Instead, anchoring is likely to be operating at the time the subject is formulating a response.

Overall, viewers, without a system recommendation, liked the episode (mean = 3.22, where 3 = "Like it"), as is generally found with product ratings. However, asymmetry of the anchoring effect was observed at the low end: Providing an artificial low recommendation reduced this preference more so than providing a high recommendation increased the preference. This effect is explored further in Study 2.

4. STUDY 2: IMPACT OF ACTUAL RECOMMENDATIONS

Study 2 follows up Study 1 by replacing the artificially fixed anchors with actual personalized recommendations provided by a well-known and commonly used recommendation algorithm. Using the user preferences for TV shows collected in Study 1, a recommender system was designed to estimate preferences of subjects in Study 2 for unrated shows. Because participants provide input ratings before being shown any recommendations or other potential anchors, the ratings were unbiased inputs for our own recommendation system. Using a parallel design to Study 1, we examine the Anchoring Hypothesis with a recommender system comparable to the ones employed in practice online.

4.1. Methods

197 people completed the study. They were solicited from the same paid subject pool as used for Study 1 with no overlap between the subjects in the two studies. Participants received a fixed fee upon completion of the study.

In Study 2, the anchors received by subjects were based on the recommendations of a true recommender system (discussed below). Each subject watched a show that he/she had indicated not having seen before – that was recommended by an actual real-time system based on the subject's individual ratings. Since there was no significant difference observed between subjects receiving system recommendations before or after viewing a show in Study 1, all subjects in the treatment groups for Study 2 saw the system-provided rating before viewing.

Three levels were used for the recommender system's rating provided to subjects in Study 2: Low (i.e., adjusted to be 1.5 points below the system's predicted rating), Accurate (the system's actual predicted rating), and High (1.5 points above the system's predicted rating). High and Low conditions were included to learn more about the asymmetry effect observed in Study 1. In addition to the three treatment groups, a control group was included for which no system recommendation was provided. The numbers of participants in the four conditions of the study are shown in Table 4 (Section 4.2).

Based on the TV show rating data collected in Study 1, an online system was built for making TV show recommendations in real time. We compared seven popular recommendation techniques 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 approaches and their extensions [2,4,24], as well as a model-based matrix factorization algorithm [11,16] popularized by the recent Netflix prize competition [3]. Each technique was evaluated using 10-fold cross validation based on the standard mean absolute error (MAE) and coverage metrics. Although the performances are comparable, the item-based CF performed slightly better than other techniques (measured in predictive accuracy and coverage). Also because the similarities between items could be pre-computed, the item-based technique performed much faster than other techniques. Therefore the standard item-based collaborative filtering approach was selected for our recommender system.

During the experiments, the system took as input subject's ratings of shows that had been seen before or for which the participant had indicated familiarity. In real time, the system predicted ratings for all unseen shows and recommended one of the unseen shows for viewing. To avoid possible show effects (e.g., to avoid selecting shows that receive universally bad or good predictions) as well as to assure that the manipulated ratings (1.5 points above/below the predicted rating) could still fit into the 5-point rating scale, only shows with predicted rating scores between 2.5 and 3.5 were recommended. When making recommendations, the system examined each genre in alphabetical order (i.e., comedy first, followed by drama, mystery, reality, and sci-fi) and went through all unseen shows within each genre alphabetically until one show with a predicted rating between 2.5 and 3.5 was found. This show was then recommended to the subject. When no show was eligible for recommendation, subjects were automatically re-assigned to one of the treatment groups in Study 1.

Our TV show recommender system made suggestions from a list of the 105 most popular TV shows that have aired in the recent decade according to a ranking posted on TV.com. Among the 105 shows, 31 were available for online streaming on Hulu.com at the time of the study and were used as the pool of shows recommended to subjects for viewing. Since our respondents rated shows, but viewed only a single episode of a show, we needed a procedure to select the specific episode of a show for viewing. For each available show, we manually compared all available episodes and selected the episode that received a median aggregated rating by Hulu.com users to include in the study. This procedure maximized the representativeness of the episode for each show, avoiding the selection of outlying best or worst episodes that might bias the participant's rating. Table 3 shows the distributions of rated and viewing-available shows by genre.

The procedure was largely identical to the Before and Control conditions used for Study 1. However, in Study 2, as indicated earlier, subjects did not all view the same show. TV episodes were again streamed from Hulu.com. The episode watched was either approximately 22 or 45 minutes in duration. For all subjects, the viewing timer was set at 20 minutes, as in Study 1. Subjects were instructed that they would not be able to proceed until the timer reached zero; at which time they could choose to stop and proceed to the next part of the study or to watch the remainder of the episode before proceeding.

Table 3. Distribution of Shows.

Genre	Number of Shows	Available for Viewing
Comedy	22	7
Drama	26	8
Mystery/Suspense	25	4
Reality	15	4
Sci Fi and Fantasy	17	8
Total	105	31

4.2. Results

Since the subjects did not all see the same show, the preference ratings for the viewed show were adjusted for the predicted ratings of the system, 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:

Rating Drift = Actual Rating – Predicted Rating.

Predicted Rating represents the rating of the TV show watched by the user during the study as predicted by the recommendation algorithm (before any perturbations to the rating are applied), and Actual Rating is the user's rating value for this TV show after watching the episode. Therefore, positive/negative Rating Drift values represent situations where the user's submitted rating was higher/lower than the system's rating, as possibly affected by positive/ negative perturbations (i.e., high/low anchors).

Similarly to Study 1, our preliminary analyses using general linear models indicated that none of the variables collected in the survey (such as demographics, etc.) demonstrated significance in explaining the response variable. The mean (standard deviation) values across the four conditions of the study for this variable are shown in Table 4. Using a one-way ANOVA, overall the three experimental groups (i.e., High, Low, and Accurate) significantly differed ($F(2, 147) = 3.43, p = .035$).

Table 4. Mean (SD) Rating Drift of the Viewed TV Show by Experimental Condition, Study 2.

	Study 2	
Group	N	Mean (SD)
High	51	0.40 (1.00)
Control	48	0.14 (0.94)
Accurate	51	0.13 (0.96)
Low	47	-0.12 (0.94)

Providing an accurate recommendation did not significantly affect preferences for the show, as compared to the Control condition (two-tailed $t(97) = 0.023, p = .982$). Consistent with Study 1, the High recommendation condition led to inflated ratings compared to the Low condition (one-tailed $t(96) = 2.629, p = .005$). The effect size was of slightly less magnitude with Cohen's $d = 0.53$, a medium effect size. However, unlike in Study 1, the anchoring effect in Study 2 is symmetric at the High and Low end. There was a marginally significant effect of the recommendation being lowered compared to being accurate ($t(96) = 1.305, p = .098$, Cohen's $d = .30$), and a marginally significant effect at the High end compared to receiving Accurate recommendations ($t(100) = 1.366, p = .088$, Cohen's $d = .23$). Similar effects are observed when comparing High/Low to Control condition. In summary, the Anchoring Hypothesis is supported in Study 2, consistently with Study 1. However, the anchoring effects were symmetric in the overall analysis of Study 2 at the High and Low ends.

To pursue the results further, we recognize that one source of variation in Study 2 as compared to Study 1 is that different shows were observed by the subjects. As it turns out, 102 of the 198 subjects in Study 2 (52%) ended up watching the same Comedy show. As a result, we are able to perform post-hoc analyses, paralleling the main analyses, limited to this subset of viewers. The mean (standard deviation) values across the four conditions of these subjects for the main response variable are shown in Table 5. Using the same response variable of rating drift, the overall effect across the experimental conditions was marginally maintained ($F(2, 77) = 2.70, p = .07$. Providing an accurate recommendation still did not significantly affect preferences for the show, as compared to the Control condition (two-tailed $t(47) = 0.671, p = .506$). Consistent with Study 1 and the overall analyses, the High recommendation condition led to inflated ratings compared to the Low condition (one-tailed $t(51) = 2.213, p = .016$). The effect size was also comparable to the overall effect magnitude with Cohen's $d = 0.61$, a medium effect size.

However, for the limited sample of subjects who watched the same episode, the effects at the High and Low end were not symmetric. Compared to receiving an Accurate recommendation, there was a significant effect of the recommendation being raised ($t(52) = 1.847, p = .035$, Cohen's $d = .50$), but not of being lowered ($t(51) = 0.286, p = .388$).

Table 5. Mean(SD) Rating Drift for Subjects Who Watched the Same Comedy Show in Study 2.

Group	N	Mean (SD)
High	27	0.81 (0.82)
Control	22	0.53 (0.76)
Accurate	27	0.37 (0.93)
Low	26	0.30 (0.86)

Thus, the indicated asymmetry of the anchoring effect is different from the asymmetry present in Study 1, being at the High end rather than the Low end. Also, the asymmetry is not robust across the overall data. Indicated is that the underlying cause of asymmetries is situational, in this case depending upon specific TV show effects. When looking at effects across different TV shows (Table 4), the show effects average out and symmetry is observed overall. When looking at effects for a particular show (Tables 2 and 5), idiosyncratic asymmetries can arise.

5. STUDY 3: ACTUAL RECOMMENDATIONS WITH JOKES

Study 3 provides a generalization of Study 2 within a different content domain, applying a recommender system to joke preferences rather than TV show preferences. As in Study 2, the procedure uses actual recommendations provided by a commonly used recommendation algorithm. A within-subjects design also allows us to investigate behavior at an individual level of analysis, rather than in the aggregate. We apply a wider variety of perturbations to the actual recommendations for each subject, ranging from -1.5 to 1.5, the values used in Study 2, rather than just using a single perturbation per subject.

5.1. Methods

61 people received a fixed fee for completing the study. They were solicited from the same paid subject pool used for Studies 1 and 2 with no overlap across the three studies.

As with Study 2, the anchors received by subjects were based on the recommendations of a true recommender system. The item-based collaborative filtering technique was used to maintain consistency with Study 2. The same list of 100 jokes was used during the study, though the order of the jokes was randomized between subjects. The jokes and the rating data for training the recommendation algorithm were taken from the Jester Online Joke Recommender System repository [12]. Specifically, we used their Dataset 2, which contains 150 jokes. To get to our list of 100, we removed those jokes that were suggested for removal at the Jester website (because they were either included in the "gauge set" in the original Jester joke recommender system or because they were never displayed or rated), jokes that more than one of the coauthors of our study identified as having overly objectionable content, and finally those jokes that were greatest in length (based on word count).

The procedure paralleled that used for Study 2 with changes adapted to the new context. Subjects first evaluated 50 jokes, randomly selected and ordered from the list of 100, as a basis for providing recommendations. The same 5-star rating scale with half-star ratings from Studies 1 and 2 was used, affording a 9-point scale for responses. Next, the subjects received 40 jokes with a predicted rating displayed. Thirty of these predicted ratings were perturbed, 5 each using perturbations of -1.5, -1.0, -0.5, +0.5, +1.0, and +1.5. The 30 jokes that were perturbed were determined pseudo-randomly to assure that the manipulated ratings would fit into the 5-point rating scale. First, 10 jokes 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, 10 jokes with predicted rating scores between 2.0 and 4.0 were selected randomly to receive perturbations of -1.0 and +1.0. Then, 10 jokes with predicted rating scores between 1.5 and 4.5 were selected randomly to receive perturbations of -0.5 and +0.5. Ten predicted ratings were not perturbed, and were displayed exactly as predicted. These 40 jokes were randomly intermixed. Following the first experimental session (3 sessions were used in total), the final 10 jokes were added as a control. A display was added on which subjects provided preference ratings for the 10 jokes with no predicted rating provided, again in random order. Finally in all sessions, subjects completed a short demographic survey.

5.2. Results

As with Study 2, the main response variable for Study 3 was Rating Drift (i.e., Actual Rating – Predicted Rating). As an illustration of the overall picture, Figure 2 shows the mean Rating Drift, aggregated across items and subjects, for each perturbation used in the study. In the aggregate, there is a linear relationship both for negative and positive perturbations. For comparison purposes, Table 6 shows the mean (standard deviation) values across the four perturbation conditions of Study 3 that were comparable to those used in Study 2 (aggregating across all relevant Study 3 responses). The general pattern for Study 3—using jokes and within-subjects design—parallels that for Study 2—using TV shows and a between-subjects design.

The within-subjects design also allows for analyses of the Anchoring Hypothesis at the individual level. We began by testing the slopes across subjects between negative and positive perturbations, and no significant difference was observed (t(60) = 1.39, two-tailed p = .17). We also checked for curvilinearity for each individual subject for both positive and negative perturbations. No significant departures from linearity were observed, so all reported analyses use only first-order effects. As an indicator of the magnitude of the effect, we examined the distribution of the correlation coefficients for the individual analyses. The mean magnitude of the relationship is 0.37, with values ranging from -0.27 to 0.87.

Overall, the analyses strongly suggest that the effect of perturbations on rating drift is not discrete. Perturbations have a continuous effect upon ratings with, on average, a drift of 0.35 rating points occurring for every rating point of perturbation (e.g., mean rating drift is 0.53 for a perturbation of +1.5).

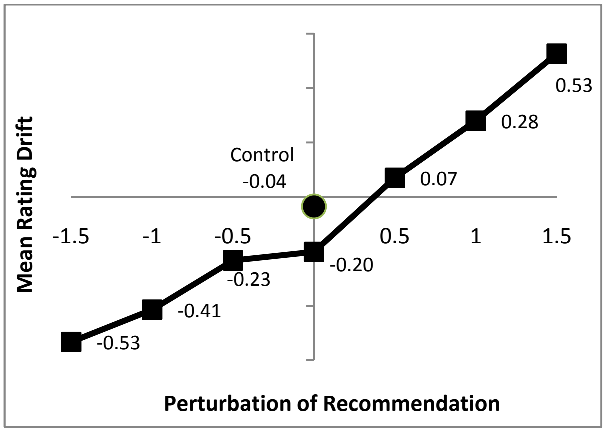

Figure 2. Mean Rating Drift as a Function of the Amount of Rating Perturbation and for Control Condition in Study 3.

Table 6. Mean (SD) Rating Drift, in the Comparable Conditions Used in Study 2 (±1.5, 0, Control), for Study 3.

Group	N	Mean (SD)
High	305	0.53 (0.94)
Control	320	-0.04 (1.07)
Accurate	610	-0.20 (0.97)
Low	305	-0.53 (0.95)

6. DISCUSSION AND CONCLUSIONS

We conducted three laboratory experiments and systematically examined the impact of recommendations on consumer preferences. The research integrates ideas from behavioral decision theory and recommender systems, both from practical and theoretical standpoints. The results provide strong evidence that biased output from recommender systems can significantly influence the preference ratings of consumers.

From a practical perspective, the findings have several important implications. First, they suggest that standard performance metrics for recommender systems may need to be rethought to account for these phenomena. If recommendations can influence consumer-reported ratings, then how should recommender systems be objectively evaluated? Second, how does this influence impact the inputs to recommender systems? If two consumers provide the same rating, but based on different initial recommendations, do their preferences really match in identifying future recommendations? Consideration of issues like these arises as a needed area of study. Third, our findings bring to light the potential impact of recommender systems on strategic practices. If consumer choices are significantly influenced by recommendations, regardless of accuracy, then the potential arises for unscrupulous business practices. For example, it is well-known that Netflix uses its recommender system as a means of inventory management, filtering recommendations based on the availability of items [26]. Taking this one step further, online retailers could potentially use preference bias based on recommendations to increase sales.

Further research is clearly needed to understand the effects of recommender systems on consumer preferences and behavior. Issues of trust, decision bias, and preference realization appear to be intricately linked in the context of recommendations in online marketplaces. Additionally, the situation-dependent asymmetry of these effects must be explored to understand what situational characteristics have the largest influence. Moreover, future research is needed to investigate the error compounding issue of anchoring: How far can people be pulled in their preferences if a recommender system keeps providing biased recommendations? Finally, this study has brought to light a potentially significant issue in the design and implementation of recommender systems. Since recommender systems rely on preference inputs from users, bias in these inputs may have a cascading error effect on the performance of recommender system algorithms. Further research on the full impact of these biases is clearly warranted.

7. ACKNOWLEDGMENT

This work is supported in part by the National Science Foundation grant IIS-0546443.

REFERENCES

[1] Adomavicius, G., and Tuzhilin, A. 2005. "Toward the Next Generation of Recommendation System: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Transactions on Knowledge and Data Engineering*, 17, (6), 734-749.

[2] Bell, R.M., and Koren, Y. 2007. "Improved Neighborhood-Based Collaborative Filtering," *KDD Cup'07*, San Jose, California, USA, 7-14.

[3] Bennett, J., and Lanning, S. 2007. "The Netflix Prize," *KDD-Cup and Workshop*, San Jose, CA, www.netflixprize.com.

[4] Breese, J.S., Heckerman, D., and Kadie, C. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *14th Conf. on Uncertainty in Artificial Intelligence*, Madison, WI.

[5] Chapman, G., and Bornstein, B. 1996. "The More You Ask for, the More You Get: Anchoring in Personal Injury Verdicts," *Applied Cognitive Psychology*, 10, 519-540.

[6] Chapman, G., and Johnson, E. 2002. "Incorporating the Irrelevant: Anchors in Judgments of Belief and Value.," in *Heuristics and Biases: The Psychology of Intuitive Judgment,* T. Gilovich, D. Griffin and D. Kahneman (eds.). Cambridge: Cambridge University Press, 120-138.

[7] 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.

[8] Deshpande, M., and Karypis, G. 2004. "Item-Based Top-N Recommendation Algorithms," *ACM Trans. Information Systems*, 22, (1), 143-177.

[9] Epley, N., and Gilovich, T. 2010. "Anchoring Unbound," *J. of Consumer Psych*, 20, 20-24.

[10] Flynn, L.J. January 23, 2006. "Like This? You'll Hate That. (Not All Web Recommendations Are Welcome.)." *New York Times*, from http://www.nytimes.com/2006/01/23/technology/23recommend.html.

[11] Funk, S. 2006. "Netflix Update: Try This at Home." 2010, from http://sifter.org/~simon/journal/20061211.html.

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

[13] Jacowitz, K.E., and Kahneman, D. 1995. "Measures of Anchoring in Estimation Tasks," *Personality and Social Psychology Bulletin*, 21, 1161-1166.

[14] Johnson, J.E.V., Schnytzer, A., and Liu, S. 2009. "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, 410-434.

[15] Komiak, S., and Benbasat, I. 2006. "The Effects of Personalization and Familiarity on Trust and Adoption of Recommendation Agents," *MIS Quarterly*, 30, (4), 941-960.

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

[17] Ku, G., Galinsky, A.D., and Murnighan, J.K. 2006. "Starting Low but Ending High: A Reversal of the Anchoring Effect in Auctions," *J. of Personality and Social Psych*, 90, 975-986.

[18] Lichtenstein, S., and Slovic, P. (eds.). 2006. *The Construction of Preference*. Cambridge: Cambridge University Press.

[19] Mcnee, S.M., Lam, S.K., Konstan, J.A., and Riedl, J. 2003. "Interfaces for Eliciting New User Preferences in Recommender Systems," in *User Modeling 2003, Proceedings*. Berlin: Springer-Verlag Berlin, 178-187.

[20] Mussweiler, T., and Strack, F. 2000. "Numeric Judgments under Uncertainty: The Role of Knowledge in Anchoring," *Journal of Experimental Social Psychology*, 36, 495-518.

[21] Northcraft, G., and Neale, M. 1987. "Experts, Amateurs, and Real Estate: An Anchoring-and-Adjustment Perspective on Property Pricing Decisions," *Organizational Behavior and Human Decision Processes*, 39, 84-97.

[22] Pu, P., and Chen, L. 2007. "Trust-Inspiring Explanation Interfaces for Recommender Systems," *Knowledge-Based Systems*, 20, (6), Aug, 542-556.

[23] Russo, J.E. 2010. "Understanding the Effect of a Numerical Anchor," *Journal of Consumer Psychology*, 20, 25-27.

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

[25] Schonfeld, E. July 2007. "Click Here for the Upsell." *CNNMoney.com*, from http://money.cnn.com/magazines/business2/business2_archive/2007/07/01/100117056/index.htm.

[26] Shih, W., S., K., and Spinola, D. 2007. "Netflix," *Harvard Business School Publishing*, (case number 9-607-138).

[27] Strack, F., and Mussweiler, T. 1997. "Explaining the Enigmatic Anchoring Effect: Mechanisms of Selective Accessibility," *Journal of Personality and Social Psychology*, 73, 437-446.

[28] Swearingen, K., and Sinha, R. 2001. "Beyond Algorithms: An Hci Perspective on Recommender Systems," *ACM SIGIR 2001 Workshop on Recommender Systems*, New Orleans, Louisiana.

[29] Thorsteinson, T., Breier, J., Atwell, A., Hamilton, C., and Privette, M. 2008. "Anchoring Effects on Performance Judgments," *Organizational Behavior and Human Decision Processes*, 107, 29-40.

[30] Tversky, A., and Kahneman, D. 1974. "Judgment under Uncertainty: Heuristics and Biases," *Science*, 185, 1124-1131.