Information Systems Research

Publication details, including instructions for authors and subscription information:
http://pubsonline.informs.org

Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects

Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, Jingjing Zhang

To cite this article:
Gediminas Adomavicius, Jesse C. Bockstedt, Shawn P. Curley, Jingjing Zhang (2013) Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects. Information Systems Research 24(4):956-975. https://doi.org/10.1287/isre.2013.0497

Information Systems Research
Vol. 24, No. 4, December 2013, pp. 956–975
ISSN 1047-7047 (print) | ISSN 1526-5536 (online)

http://dx.doi.org/10.1287/isre.2013.0497

Do Recommender Systems Manipulate Consumer Preferences? A Study of Anchoring Effects

Gediminas Adomavicius
Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455,
gedas@umn.edu

Jesse C. Bockstedt
Management Information Systems, Eller College of Management, The University of Arizona, Tucson, Arizona 85721,
bockstedt@email.arizona.edu

Shawn P. Curley
Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455,
curley@umn.edu

Jingjing Zhang
Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, Indiana 47405,
jjzhang@indiana.edu

Recommender systems are becoming a salient part of many e-commerce websites. Much research has focused on advancing recommendation technologies to improve accuracy of predictions, although behavioral aspects of using recommender systems are often overlooked. In our studies, 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 across a variety of conditions, which were surveyed immediately following program viewing. Study 3 investigated the granularity of the observed effects within individual participants. Results provide strong evidence that the rating presented by a recommender system serves as an anchor for the consumer's constructed preference. Viewers' preference ratings are malleable and can be significantly influenced by the recommendation received. The effect is sensitive to the perceived reliability of a recommender system and, thus, not a purely numerical or priming-based effect. Finally, the effect of anchoring is continuous and linear, operating over a range of perturbations of the system. These general findings have a number of important implications (e.g., on recommender systems performance metrics and design, preference bias, potential strategic behavior, and trust), which are discussed.

Key words: anchoring effects; behavioral decision making; behavioral economics; electronic commerce; experimental research; preferences; recommender systems
History: Michael Smith, Senior Editor; Gautam Pant, Associate Editor. This paper was received on April 30, 2011, and was with the authors 14 months for 3 revisions. Published online in *Articles in Advance* September 5, 2013.

1. Introduction

Recommender systems are important decision aids in the electronic marketplace and an integral part of the business models of many firms. Such systems provide product suggestions to consumers, allowing firms to better serve their customers and increase sales. For example, it has been reported that a recommender system could account for 10%–30% of an online retailer's sales (Schonfeld 2007) and that roughly two-thirds of the movies rented on Netflix were suggested by its recommender system (Flynn 2006). Research in the area of recommender systems has focused almost exclusively on the development and improvement of the algorithms for making accurate recommendations and predictions. Less well studied are the behavioral implications of using recommender systems in the electronic marketplace.

Most recommender systems use consumer ratings of experienced items as inputs for the system's computational techniques (based on methodologies from statistics, data mining, or machine learning), which estimate preferences for items that have not yet been consumed by the individual. These estimated preferences are often presented in the form of predicted "system ratings," which indicate an expectation of how well the consumer will like an item, serving as recommendations. The subsequent consumer ratings (provided after item consumption) complete a feedback loop that is central to a recommender system's use and value, as illustrated in Figure 1. 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 actual ratings of the users. In three 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. In study 1 we test the effects of artificial system ratings on consumer preferences, and in study 2 we examine the effects of actual system recommendations that have been perturbed. Furthermore, study 3 investigates the granularity of the observed effects using a within-subjects design, and examines these effects in a different setting, using jokes instead of TV shows.

Providing consumers with a predicted "system rating" can potentially introduce anchoring biases and significantly influence their subsequent rating of an item. Our studies are designed to test for the existence of this influence, to examine its magnitude, and to investigate potential explanations of any effects. As noted by Cosley et al. (2003), biases in the ratings provided by users can lead to three potential problems: (i) biases can contaminate the recommender system's inputs, 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 to operate in their favor.

Researchers in behavioral decision making, behavioral economics, and applied psychology have found that people are often influenced by elements in the environment in which preferences are constructed (e.g., Chapman and Johnson 2002, Lichtenstein and Slovic 2006, Tversky and Kahneman 1974). Our objective is to understand the influence of recommender systems' predicted ratings on consumers' preferences. In particular, we explore four issues related to the impact of recommender systems: (1) The *anchoring* issue—are people's preference ratings for items they just consumed drawn toward predictions that are given to them? Understanding any potential anchoring effect, particularly at the point of consumption, is the principal goal of this study. (2) The *timing* issue—does it matter whether the system's prediction is presented before or after user's consumption of the item? One posited explanation for anchoring effects is that showing the prediction prior to consumption provides a prime that influences the user's consumption experience and his or her subsequent rating of the consumed item. If priming is a factor, then 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 is evidence that anchoring in recommender systems is not a purely numeric effect. (4) The *granularity* issue—what is the anchoring effect size and functional form with respect to a continuum of system-generated rating predictions?

Understanding the impact of recommendations on user preferences provides opportunities to further improve recommender systems. For example, consider the fact that the best performing research teams in the recent $1M Netflix Prize competition (www.netflixprize.com), using the latest machine learning developments in recommendation algorithms (Koren 2009, Piotte and Chabbert 2009, Töscher et al. 2009), achieved recommendation accuracy of around 0.85 (as measured in root mean square error) on a scale of length 4.0 (i.e., from one to five stars), leaving a lot of room for improvement. This paper provides a promising avenue for exploring potential further performance improvements, first, by establishing that the user's reported preference ratings can be impacted just by receiving a system recommendation and, second, by measuring the nature and magnitude of this effect. Future work can use these results as a basis for new directions in designing recommendation algorithms and user interfaces that are able to improve recommendation performance by compensating for the anchoring effects.

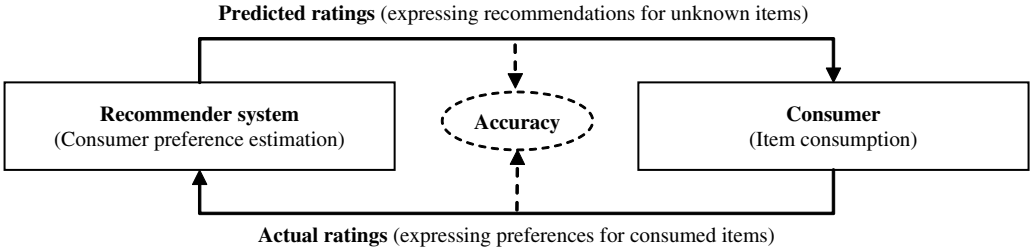
2. Background and Hypotheses

2.1. Anchoring Effects in Judgment

Behavioral decision-making research has shown 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 in applying an anchoring-and-adjustment heuristic (Tversky and Kahneman 1974, Chapman and Johnson 2002), 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 (2010) identified three waves of research on anchoring. The first wave, beginning with the initial study by Tversky and Kahneman (1974), focused on establishing anchoring and adjustment as leading to biases in judgment. This research primarily used almanac questions (the answers for which can be looked up in a reference source) in an artificial survey setting; however, the effect has proved robust with respect to context, e.g.,

Figure 1 Ratings as Part of a Feedback Loop in Consumer-Recommender Interactions



anchoring is found with mock jurors making legal case decisions (Chapman and Bornstein 1996), with students assessing instructors (Thorsteinson et al. 2008), and with real estate professionals making house appraisals (Northcraft and Neale 1987). This first research wave is mature and seems to have run its course. The second and third waves (within which our research falls) are still active research areas, however.

The second wave focuses on psychological explanations for anchoring effects. Russo (2010) identified three types of explanations from the literature. The first explanation argues 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 (e.g., Jacowitz and Kahneman 1995). When preferences are recalled from past experience, there is often uncertainty about one's preferences and this mechanism can be operative. However, in our research, preference ratings are collected at the time of consumption when uncertainty is highly minimized. A second explanation is that the anchor leads to biased retrieval of anchor-consistent knowledge. Again, when preferences are recalled and one's preference ratings arise from a process of reasoning from prior knowledge, this explanation might pertain. However, in our studies, preference ratings are made at the time of consumption and so this explanation is expected to be inoperative, as well. A third explanation is that of numerical priming. This explanation is plausible in our context and connects to the timing issue, as described in §1. To these three explanations we would add a possibility, connected to the system reliability issue, that the anchor is viewed as a suggestion provided by the context as to the correct answer (e.g., Chapman and Bornstein 1996, Epley and Gilovich 2010, Northcraft and Neal 1987). In preference construction, the user's trust in the recommender system's estimate may serve as a measure of the degree of belief that the system's recommendation can be taken as a "correct" value in a predictive sense. As potentially operative in our setting, study 1 is designed to address directly these latter two potential explanations of anchoring effects.

The third wave of anchoring research 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" (Epley and Gilovich 2010, p. 21). Some recent examples of the third wave are works by Johnson et al. (2009), who study anchoring in a real-world setting of horserace betting, and Ku et al. (2006), who investigate anchoring effects in auctions.

In this framework, our studies are designed to make central contributions within the less-studied second and third waves of research on anchoring in several important ways. Foremost, we study anchoring within the context of recommender systems. This is an area of great practical interest that has only received attention in a single study by Cosley et al. (2003). Our work significantly extends their initial investigation as detailed in §2.3. Theoretically, relative to the more general research on anchoring effects, the focus on recommender systems contributes by addressing the effects (a) in a preference setting (rare in the literature; see §2.2), and (b) at the time of consumption (singular in the literature, as discussed more fully in §§2.2 and 2.3). Finally, within the context of recommender systems, we provide evidence relative to the proposed explanations for anchoring effects. We directly test two of the proposed explanations described above: the possibility that the anchor serves in priming the preference, and the explanation that the anchor provides content that is relevant to the user's preference. By comparison with the study by Cosley et al. (2003), our studies also provide indirect evidence relative to the other two proposed explanations that have been suggested for anchoring effects: the role of uncertainty in driving the effect, and the suggestion that the anchor leads to a biased retrieval process. We expand on all of these contributions in the next two subsections.

2.2. Anchoring Effects on Preference Ratings

The effect of anchoring on preference construction is an important open issue. For example, Ariely et al. (2003, p. 75) state that "the vast majority of anchoring experiments in the psychological literature have focused on how anchoring corrupts subjective judgment, not subjective valuation or preference." Past

studies have largely used tasks for which there is a verifiable outcome, leading to a bias measured against an objective performance standard (e.g., review by Chapman and Johnson 2002). In the recommendation setting, the judgment is a subjective preference and is not verifiable against an objective standard.

One of the few papers identified in the mainstream anchoring literature that has looked directly at anchoring effects in preference construction is that of Schkade and Johnson (1989). However, their work studied preferences between abstract, stylized, simple (two-outcome) lotteries. This preference situation is far removed from the more realistic situation that we address in our work. More similar to our setting, Ariely et al. (2003) observed anchoring in bids provided by students in a classroom setting participating in auctions of consumer products (e.g., wine, books, chocolates). However, their design did not allow bidders to experience the items and, thus, preferences were based on expectations of experience, not on actual experience or immediate consumption. Our research, instead, focuses on preferences immediately following consumption when we expect them to be newly formed and clearest, a point the significance of which is detailed in the next section.

2.3. Anchoring Effects with Recommender Systems

Very little research has explored how the cues provided by recommender systems influence online consumer behavior. The work that comes closest to ours is by Cosley et al. (2003), which deals with a related but significantly different anchoring phenomenon in the context of recommender systems. Cosley et al. explored the effects of system-generated recommendations on user reratings of movies. They found that users showed high test-retest consistency when being asked to rerate a movie with no prediction provided. However, when users were asked to rerate a movie while being shown a "predicted" rating that was altered upward or downward from their original rating for the movie by a single fixed amount of one rating point (providing a high or a low anchor), users tended to give higher or lower ratings, respectively (compared to a control group receiving accurate original ratings).

The comparison and contrast between their work and ours is summarized in Table 1. The research setting of users acting with recommender systems is similar to that of Cosley et al. (2003), and, therefore, allows us to build upon and extend their research. As noted, one distinguishing and interesting feature of this setting is that the participant is performing a *preference* task. The potential role and operation of anchoring within this type of task, here in the context of recommender systems, provides an important extension of the anchoring literature more generally (e.g., as called for by Ariely et al. 2003).

Table 1 Comparison and Contrast Between Cosley et al. (2003) Study and Our Reported Studies

	Cosley et al. (2003)	Current studies
Setting	Recommender systems	Recommender systems
Type of task	Preference (no objective standard)	Preference (no objective standard)
Stimuli	Multiple movies	Single TV shows, multiple TV shows, multiple jokes
Recommendations	System-based	System-based, plus artificially generated
Manipulations	Two: High vs. Low	Multiple: High vs. Low; also range of manipulations
Timing (process implications)	Retrospective (retrieval; uncertainty)	Point of consumption (integrating and responding; no uncertainty)
Explanations	None	Directly (timing, perceived reliability hypotheses) and indirectly provide evidence relative to 4 possible explanations that have been posited for anchoring

There are several ways in which our work provides significant extensions to prior research. These differences provide convergent validation for the primary findings of Cosley et al. (2003) with respect to anchoring over a greater range of experimental setups. One simple extension is that we broaden the results to other context areas, beyond the effects of anchoring on aggregate judgments across different movies; specifically, we examine effects across different television shows and across a variety of jokes. Further, our designs allow us to focus attention on the effects for a single television show. As we will see, this helps us to better understand the asymmetry of anchoring effects in the current preference-oriented context of using recommender systems.

More importantly, we use a broader range of recommendations, not only manipulating the system's rating to a higher or lower recommendation, but also using artificially high and low recommendations that are not tied to the system's rating. Again, this provides a degree of convergent validity to the research. More significantly, we also employ a design in study 3 that uses a *continuum* of perturbations, not just a single high and a single low manipulation. The within-subjects design of study 3, with each user seeing system's recommendations that are manipulated at different levels, provides the first evidence of the functional form of the anchoring effect at the level of each individual user.

Aside from these offerings, another significant contribution of our research lies in its implications for

the users' decision processes that potentially underlie anchoring effects. As noted by Chapman and Johnson (2002), anchors can impact the decision-making process in three stages: (1) retrieval and selection of information, (2) integration of information, and (3) formation of a response. Cosley et al. (2003) focused upon the impact of recommendation anchors in the first stage: retrieval and selection of information. Users were asked to remember how well they liked movies from their past experiences, where their original preference construction may have occurred months or years prior to the study. In contrast, our studies investigate the effects of anchors at the time of consumption, thereby focusing on the second and third stages of the decision-making process, i.e., integrating information and formulating a response.

From a process standpoint, this difference is very significant. As noted in §2.1, the behavioral decision-making literature has suggested *uncertainty* and *biased recall* as two plausible explanations for the occurrence of anchoring effects in judgment. Cosley et al.(2003) established anchoring effects on preferences in recall tasks, i.e., tasks in which these explanations may also be operable. However, we measure the immediate effects of anchoring, when preferences are expressed at the time of consumption. Uncertainty and biased recall do not apply in these processes that highlight the formation of preference. Thus, our study provides evidence relevant to the operation of these mechanisms in preference settings, providing an important extension to the earlier research.

More directly, through experimental manipulation of independent variables, we also test the other two explanations, in a preference situation, that have been suggested in other judgmental settings. In study 1, we investigate the *priming* explanation by varying the timing of the recommendation, and, we investigate the explanation that the recommendation is understood as *providing content relevant to one's preference* by varying the reliability of the recommendation.

2.4. Hypotheses

Turning to the explicit hypotheses, since anchoring has been observed in other settings, though different than ours, we begin with the conjecture that the rating provided by a recommender system serves as an anchor for the consumer's constructed preference. Insufficient adjustment away from the anchor will lead to a preference rating that is observably 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.*

As noted earlier, one explanation for anchoring biases is that the anchor can serve as a prime for upcoming experience. We employ the standard reasoning for exploring the potential of a priming effect: If this dynamic operates in the current setting, then receiving the recommendation prior to consumption should lead to greater anchoring effects than receiving the recommendation after consumption.

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*.

Also noted earlier, another proposed explanation is that the user perceives the anchor as providing evidence as to a correct answer in situations where an objective standard exists (e.g., Mussweiler and Strack 2000). When applied to the use of recommender systems and preferences, the explanation might surface instead as an issue of the consumer's trust in the system. Using a laboratory experiment involving commercial recommender systems, Komiak and Benbasat (2006) found that increasing cognitive and emotional trust improved consumer's intentions to accept the recommendations. Research has also highlighted the potential role of human-computer interaction and system interface design in achieving high consumer trust and acceptance of recommendations (e.g., McNee et al. 2003, Pu and Chen 2007, Swearingen and Sinha 2001, Wang and Benbasat 2007). In contrast, anchoring effects may be a purely numerical phenomenon. Anchoring effects have been observed with arbitrary anchors (e.g., Tversky and Kahneman 1974), and impossibly extreme anchor values (e.g., Strack and Mussweiler 1997) for almanac quantities (e.g., the number of countries in the United Nations). These studies suggest that the anchoring effect may be purely a numerical phenomenon, and that the quality of the anchor may be less important.

To the extent that the phenomenon as exhibited with preference is purely numerically driven, weakening of the recommendation should not lessen the effects of anchoring. 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. The following hypothesis captures the latter expectation:

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 system perceived as less reliable*.

Table 2 Summary of Differences Across Studies 1–3

	Study 1	Study 2	Study 3
Setting	Preference Judgments with a Recommender System		
Stimuli	Single TV show [controls for stimulus differences]	Multiple TV shows [controls for individual differences], (single TV show)	Multiple jokes
Hypotheses tested	Anchoring, timing, perceived system reliability, (asymmetry)	Anchoring, (asymmetry)	
Primary independent variables	Artificial anchors (high/low); timing (before/after); system reliability (high/low)	Perturbation of recommender system-based anchors {−1.5, 0, +1.5}	Perturbation of recommender system-based anchors {−1.5, −1, −0.5, 0, +0.5, +1, +1.5}
Purpose	(a) Build database for Study 2 (b) Testing multiple hypotheses	Test anchoring hypothesis with a real system	Model the functional form of the anchoring effect
Relative to Cosley et al. (2003)	New study	Design follows Cosley et al. (with changes noted in Table 1)	New study

A final note is that Thorsteinson et al. (2008), in their study of job performance ratings, found an asymmetry of the anchoring effect such that high anchors produced a larger effect than did low anchors. In interacting with recommender systems, people tend to have more experience with products for which they have an affinity (at the high end of the scale) than with those that they dislike. For example, YouTube recently moved away from their original star-based rating system because of the overwhelming dominance of five-star ratings.[1] However, Thorsteinson et al. (2008) provided no explanation for the asymmetry they observed. We leave open the possibility of asymmetric anchoring effects in recommender systems and incorporate it as an exploratory factor in our design; however, we provide no hypothesis as to asymmetry.

2.5. Overview of Studies

We conducted three controlled laboratory experiments, in which system predictions presented to participants were biased upward and downward to test our hypotheses. In all three studies, participants consumed an item (television shows or jokes) and immediately reported their preference rating for the item. The three studies were designed to capture preferences as they are constructed. The Anchoring Hypothesis is the primary hypothesis of interest and is investigated in all three studies. Similarities and differences across the studies are summarized in Table 2.

Study 1 was designed to demonstrate the existence of the anchoring effect. We controlled the stimulus (i.e., every participant consumed the same item) and presented participants with randomly assigned artificial system recommendations (i.e., the manipulation was not tied to the individual at all). Study 1 also incorporated factors in its design to test the Timing and Perceived System Reliability Hypotheses. Additionally, study 1 allowed us to compile a database of ratings, which was used as a training set to make targeted recommendations in study 2.

As we will observe, the results of study 1 support an effect of anchoring. The primary goal of study 2 was to test for this anchoring effect with a recommender system providing real-time personalized recommendations based on user ratings. Unlike study 1, study 2 took individual preference differences into account. Actual personalized recommendations were provided to participants by a well-known and commonly used recommendation algorithm and were then biased upward or downward to create the treatments in the study. In study 2 we controlled for contextual and item-based factors in our analysis and explored the potential for asymmetry in the anchoring effects. It provides the contrast to the study of Cosley et al. (2003) for testing indirectly the necessity of uncertainty and biased retrieval as explanations for anchoring effects in a preference task. These factors do not play a significant role in preferences at the point of consumption (our study) but could play a role in recalled preferences (their study). Thus, if anchoring effects are not observed in study 2, this would provide evidence that one or the other of these two explanations is necessary for anchoring effects to operate.

Study 3 was designed to investigate the granularity of the observed effects. The first two studies addressed the impact of large perturbations in testing the Anchoring Hypothesis, but they did not provide any insights of the functional form of the effect. In study 3 we observed the effect size over a continuum of perturbations. The within-subjects design of study 3 also allowed for analysis of individual differences among the participants, which are averaged in the between-subjects designs of studies 1 and 2. To analyze the functional relationship and explore

[1] As posted on the official YouTube blog (http://youtube-global.blogspot.com/2009/09/five-stars-dominate-ratings.html).

within-subjects effects, we needed to use a context other than television shows that allowed for multiple consumption events for each participant in a reasonable amount of time. Therefore, a side benefit of study 3 is that it allowed us to test the Anchoring Hypothesis in a different context (jokes), thereby providing evidence of the generalizability of our results.

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 versus low anchors (i.e., asymmetry); (3) to test the Timing Hypothesis and (4) to test the Perceived System Reliability Hypothesis for anchoring effects with system recommendations; and (5) to build a database of user preferences for television shows to be used in computing personalized recommendations for study 2.

3.1. Methods

3.1.1. Participants. Two-hundred and sixteen 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 people all seeing the show for the first time, left 206 responders for analysis. Participants were solicited from a paid subject pool maintained by a U.S. business school. Since there was no clear criterion to measure their performance on the task, no performance-based incentive was provided; participants received a fixed fee at the end of the study. Demographic features of the sample are summarized in the first data column of Table 3.

3.1.2. Design. In study 1 participants received *artificial* anchors, i.e., system ratings were not produced by a recommender system. All participants were shown the same TV show episode during the study and were asked to provide their rating of the show after viewing. TV shows were chosen as the subject area because they are relatively short (the episode shown was approximately 20 minutes in length), and they have a variety of features that lead to variable preferences. 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 that varied across three factors. First, the system rating was either a *low* (1.5, on a scale of 1 through 5) or *high* value (4.5). Second, the timing of the recommendation was either *before* or *after* the show was watched (but always before the viewer was asked to rate the show). Together, the first two factors form a 2×2 (high/low anchor $\times$ before/after viewing) between-subjects design (the top two rows of the design in Table 4).

Table 3 Demographic Characteristics of Participants in Studies 1–3

		Study 1		Study 2		Study 3	
Sample size		206		197		61	
Gender:	Male (%)	83	(40%)	78	(40%)	28	(46%)
	Female (%)	123	(60%)	119	(60%)	33	(54%)
Age: Mean (SD)		23.1	(7.25)	24.2	(8.30)	22.1	(5.87)
Hours of TV watched on average per week:	<1(%)	31	(15%)	38	(19%)		
	2–4 (%)	72	(35%)	64	(32%)		
	4–8(%)	69	(34%)	64	(32%)		
	12–16(%)	30	(15%)	27	(14%)		
	>20(%)	4	(2%)	4	(2%)		

Table 4 Experimental Design and Sample Sizes per Group in Study 1

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

The third factor of the system's perceived reliability intersects with this design. In the *strong* reliability conditions, instructions were (shown 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 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.

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 2×2 between-subjects design (the bottom two rows of the design in Table 4).

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 4.

3.1.3. Procedure. 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, participants 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 seven-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 two or above). Subjects rated each of these shows using a five-star scale that used verbal labels parallel to those in use by Netflix.com: * = "hate it," ** = "don't like it," *** = "like it," **** = "really like it," and ***** = "love it." Half-star ratings were also allowed, so that subjects had a nine-point scale for expressing preference. A nine-point scale for this setting was shown to have good test-retest reliability by Cosley et al. (2003). 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 of training data for the recommender system used in study 2 (to be described later).

Following this initial rating task, all participants saw the same TV episode (the pilot) of a situation comedy. A less well-known TV show was chosen to maximize the likelihood that the majority of participants were not familiar with it. The episode was streamed from Hulu.com and (including closing credits) 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 length of 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, participants were presented with a screen that provided the system recommendation with the wording appropriate to their condition (strong/weak, low/high anchor). This screen was omitted in the control condition. Following, participants rated the episode just viewed. The same five-star (nine-point) rating scale used earlier was provided for the preference rating, except that the "not able to rate" option was omitted. Finally, respondents completed a short survey that included questions on demographic information and TV viewing patterns (see Table 3).

Table 5 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)
		Control	29	3.22 (0.98)
*		After-High-Strong	28	3.43 (0.81)
*		After-Low-Strong	29	2.88 (0.79)
*	*	Before-High-Strong	31	3.48 (1.04)
*	*	Before-Low-Strong	29	2.78 (0.92)
	*	Before-High-Weak	31	3.08 (1.07)
	*	Before-Low-Weak	29	2.83 (0.75)

3.2. Results

Table 5 shows the mean of the ratings reported by participants for the viewed episode in the seven experimental groups. The experiment incorporated two intersecting 2×2 factorial experimental designs, and the first two columns of Table 5 indicate the cells involved in each of the two designs. We analyze each design in turn.

For design 1, the two manipulated factors are used to test the Anchoring and Timing Hypotheses. We begin by using a general first-order linear regression model with the reported rating of the viewed episode as the dependent variable. Since ratings were reported on a numeric interval scale with equal interval spacing (1–5 with 0.5 steps), the dependent variable was treated as continuous in our regression model and an ordinary least squares (OLS) estimator was used for ease in interpreting the results. A regression model using an ordered logit estimator to account for the ordinal nature of the dependent variable was also tested, and results qualitatively identical to the OLS regression were obtained. The reliability factor is fixed at the strong level and does not enter the analysis for design 1. The anchoring factor has two levels, either high (1) or low (0), and the timing of the recommendation is either before (0) or after (1) viewing the television episode. In addition to these two fixed factors in the model, we included age, hours of television watched, and gender (male = 1, female = 0) as controls. Subjects also rated each of the five genres identified in the study according to how much they generally liked shows within the genre, using the same five-star scale (allowing for half-stars). The responses for each of the five genres were also entered as controls in the regression model. The results are shown in Table 6.

Of the factors not central to the design, only three attained significance: the covariates age, comedy, and mystery. For age, older respondents tended to give higher ratings. Higher ratings were also given by respondents who were more favorable to comedy and mystery shows. Since the particular show used in the study was a situation comedy about a detective, these results are as expected.

Table 6 Results of the OLS Regression Models (Designs 1 and 2) of Study 1

		Design 1 ($R^2 = 0.18$)			Design 2 ($R^2 = 0.17$)		
Source		Coefficient (SE)	t statistic	Significance level	Coefficient (SE)	t statistic	Significance level
Intercept		1.068 (0.737)	1.449	0.149	1.070 (0.724)	1.476	0.142
Anchoring (High = 1)		0.492 (0.136)	3.623	$< 0.001^{***}$	0.603 (0.164)	3.689	0.000^{***}
Timing (After = 1)		0.076 (0.145)	0.522	0.602			
Reliability (Strong = 1)					0.254 (0.195)	1.299	0.196
Anchoring * Reliability					0.334 (0.280)	1.191	0.236
Gender (Male = 1)		0.044 (0.146)	0.301	0.764	0.058 (0.147)	0.396	0.693
Age		0.020 (0.010)	2.061	0.041^{*}	0.020 (0.010)	2.028	0.044^{**}
Hours watched	< 1	0.045 (0.541)	0.083	0.934	0.008 (0.541)	0.016	0.988
(Baseline: > 20)	2–4	0.102 (0.527)	0.194	0.846	−0.050 (0.525)	−0.096	0.924
	4–8	0.097 (0.530)	0.183	0.855	−0.050 (0.528)	−0.094	0.925
	12–16	0.322 (0.543)	0.593	0.554	−0.269 (0.542)	−0.496	0.620
Comedy		0.206 (0.081)	2.548	0.012^{*}	0.208 (0.081)	2.028	0.011^{*}
Drama		0.054 (0.088)	0.616	0.539	0.056 (0.088)	0.635	0.526
Mystery		0.201 (0.081)	2.490	0.014^{*}	0.190 (0.081)	2.341	0.020^{*}
Reality		0.055 (0.059)	0.926	0.356	0.063 (0.059)	1.056	0.293
Sci-Fi/Fantasy		0.011 (0.059)	0.186	0.853	0.006 (0.059)	0.097	0.923

Note. The dependent variable for each model is the rating of the viewed episode.
$^{*}p \le 0.05$, $^{**}p \le 0.01$, $^{***}p \le 0.001$.

Turning to the primary factors in design 1, firstly there is a significant observed anchoring effect of the provided artificial recommendation (Table 6, $p < 0.001$). The difference between the high and low conditions (aggregated across before and after treatments for each) was in the expected direction, showing a substantial effect between groups (one-tailed $t(115) = 3.822$, $p = 0.0001$). The corresponding regression coefficient estimate in Table 6 ($b = 0.492$) suggests that people shown a higher artificial rating reported a rating of half a star higher, on average. Using Cohen's (1988) d, which is an effect size measure used to indicate the standardized difference between the two means of the anchoring treatment groups (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.

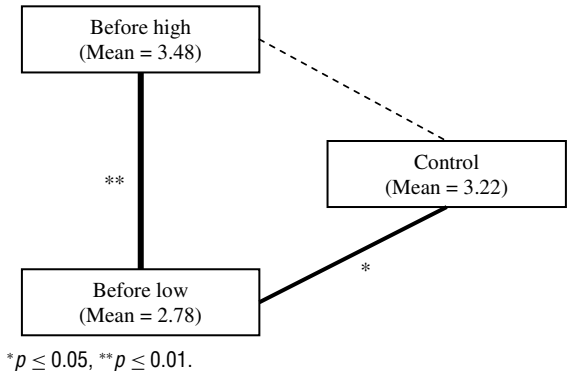
In contrast, as is apparent from Table 5 (rows marked as design 1), the Timing Hypothesis was not supported. Both the regression model (Table 6, $p = 0.602$) and the direct contrast ($t(115) = 0.043$, two-tailed $p = 0.97$, assuming equal variances) support this conclusion. This suggests that the anchoring 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 constructing a preference and formulating a response.

In design 2, we analyzed the cells indicated by the second 2×2 between-subjects design in the study (rows marked as design 2 in column 2 of Table 5) to test the Anchoring and Perceived Reliability Hypotheses. We again use a linear regression model with the reported rating of the viewed episode as the dependent variable. The timing factor is fixed at the before level and does not enter the analysis for design 2. The anchoring factor has the same two levels as with design 1 (high = 1, low = 0), and the perceived system reliability factor has two levels (strong = 1, weak = 0). Also as with design 1, the same covariates as well as gender and hours of television watched were included in the model. In addition, we note that 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. Consequently, an interaction term between timing and perceived reliability was included in the model. The results are shown in Table 6.

Of the factors not central to the design, the same three attained significance: age, comedy, and mystery. As already discussed, none of these covariates are consequential to the hypotheses of interest.

Turning to the primary factors in design 2, the Anchoring Hypothesis was again supported as in the analysis of design 1 ($b = 0.603$, $p < 0.001$). Although the Perceived Reliability Hypothesis was not supported (interaction term, $p = 0.236$), the regression model does not account for the directional nature of the hypotheses as stated. Thus, we followed up with the direct contrasts of interest. We found no significant difference between the high and low conditions with weak recommendations ($t(58) = 1.053$, $p = 0.15$), but found significant differences between the conditions with strong recommendations ($t(58) = 2.788$, $p = 0.0035$). Additionally, the overall anchoring effect was substantially reduced for the weak setting (Cohen's $d = 0.28$, small effect size

Figure 2 Contrasts of Experimental Conditions: Asymmetry of the Anchoring Effect in Study 1



range), compared to the strong recommendation setting (Cohen's $d = 0.72$, medium to large range). This provides evidence that subjects were sensitive to the perceived reliability of the recommender system, consistent with results by Komiak and Benbasat (2006) and others cited earlier. 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.[2] When an artificial high recommendation was provided (4.5), ratings were greater than those of the control group, but not significantly so ($t(58) = 0.997$, $p = 0.162$). But when an artificial low recommendation was provided (1.5), ratings were significantly lower than those of the control group ($t(56) = 1.796$, $p = 0.039$). As summarized and illustrated in Figure 2, there was an asymmetry of the effect; however, the direction was opposite to that found by Thorsteinson et al. (2008). To follow up on this difference, study 2 was designed to provide further evidence using a live recommendation system, and study 3 was designed to explore these effects at a finer level of granularity.

4. Study 2: Impact of Actual Recommendations

Study 2 followed up study 1 by replacing the artificially fixed anchors, which did not account for individuals' preference differences, with actual personalized recommendations provided by a well-known, commonly used algorithm. Study 2 thereby took individual preference differences into account and moved the test for the anchoring effect into the context of interest. With the user preferences for TV shows collected in study 1 as training data, a recommender system was designed and used to estimate preferences of participants in study 2 for unrated shows. Note that, because participants provided the input ratings before being shown any recommendations or other potential anchors, there is no reason to believe these ratings were biased inputs for our own recommendation system. Additionally, since the recommendations for study 2 were generated from a live system based on real user ratings, not all participants viewed the same television show. These changes provided a more realistic setting for the study.

4.1. Methods

4.1.1. Participants. One hundred and ninety-eight people completed the study. They were solicited from the same paid participant pool as used for study 1 with no overlap between the two studies. Participants received a fixed fee upon completion of the study. One instance was removed from the data because the participant did not take sufficient time to read the instructions and perform the required tasks, leaving 197 data points for the analysis. Demographic features of the sample are summarized in the second data column of Table 3. As observed, the samples in study 1 and study 2 are comparable along these characteristics.

4.1.2. Design. Each participant watched a show—which he or she had indicated not having seen before—that was recommended by an actual system in real time based on the participant's individual ratings. Since there was no significant difference observed between participants 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 anchor provided to participants in study 2: *low* (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). 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 8 (§4.2).

4.1.3. Recommender System and Episode Selection. Based on the TV show rating data collected in study 1, an online recommender system was built for the purpose of making TV show recommendations in real time. We compared seven popular recommendation techniques (Table 7) to find the best-performing technique for our data set. The techniques included simple user- and item-based rating average methods, user- and item-based collaborative filtering (CF) approaches and their extensions (Bell and

[2] Similar results were obtained using the after-high and after-low conditions as comparison, or using the combined high and low groups.

Table 7 Comparison of Recommendation Techniques on TV Show Rating Data Set

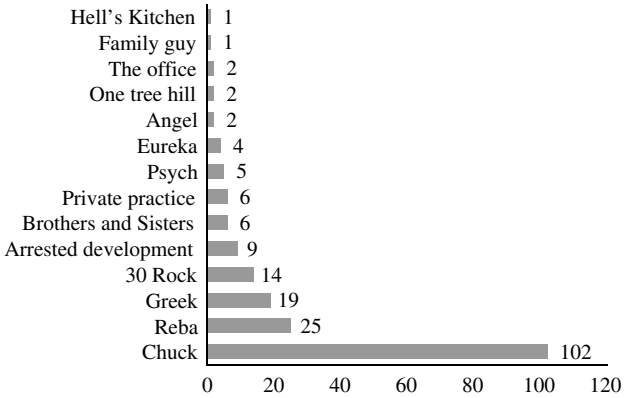
Methodology	Description	Predictive accuracy	Predictive coverage
Item average	Predicts each user-item rating as an average rating of that item (or that user, for user-based approach).	1.1550	1
User average		1.1249	1
Item-based CF	For each user-item rating to be predicted, finds most similar items that have been rated by the user (or finds other users who have similar taste, for user-based approach) and computes the weighted sum of neighbors' ratings as the predicted rating.	1.0621	0.9787
User-based CF		1.0639	0.9622
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.	1.0977	1
Item interpolation	A variation of the standard CF approach. Instead of using a similarity score to assign weights to neighbors' ratings, the algorithms learn the interpolation weights by modeling the relationships between the user and his/her neighbors through a least squares problem.	1.1073	0.9361
User interpolation		1.0937	0.8749

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 (Bennet and Lanning. 2007). Each technique was evaluated using 10-fold cross validation based on the standard mean absolute error (MAE) and coverage metrics. Collaborative filtering algorithms performed best; and, consistent with the literature (Adomavicius and Tuzhilin 2005, Deshpande and Karypis 2004, Sarwar et al. 2001), item-based CF performed slightly better than the user-based CF approach. Based on these results, we used the item-based CF approach for our recommender system.

During the experiments in study 2, the system took as input participant'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 and unfamiliar 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 five-point rating scale, only shows with predicted rating scores between 2.5 and 3.5 were recommended to participants. In the ideal case, we would have each participant watch the same show to account for show differences; however, participants have varying prior experiences and preferences, so finding a show that all participants had not seen and which had a predicted rating between 2.5 and 3.5 for each participant was not possible. However, to maximize the possibility of overlap in the show selected across participants, the system employed the following systematic TV show selection procedure. For every user, 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. When no show was eligible for recommendation, participants were automatically reassigned to one of the treatment groups in study 1.

The TV show recommender system employed in this study made suggestions from a list of the 105 most popular TV shows in the recent decade according to a popularity 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 participants for viewing. Note that our respondents rated shows, but viewed only a single episode of a show. For viewing, we selected the episode that received a median aggregated rating by Hulu.com users. 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. In the experiment, a total of 14 shows were recommended to participants, with a majority of participants (102 of 198) watching the same episode of the TV show "Chuck" (see Figure 3) because of the systematic show selection procedure mentioned above. This allowed us to perform additional analyses on this subset of subjects, where the show heterogeneity is completely removed. Alternatively, we will

Figure 3 Distribution of Shows Watched in Study



also discuss the additional controls we included in our analysis to account for show-level heterogeneity.

4.1.4. Procedure. The procedure was largely identical to the before and control conditions used for study 1. However, in study 2, as indicated earlier, participants 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 participants, the viewing timer was set at 20 minutes, as in study 1. Participants were not 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 watch the remainder of the episode before proceeding.

4.2. Results

Since the participants did not all view 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

$$\text{rating drift} = \text{actual rating} - \text{predicted rating}.$$

Predicted rating represents the rating of the TV show watched by the user as predicted by the recommendation algorithm (before any possible perturbations to the rating are applied). Actual rating is the user's reported rating for this TV show after watching the episode. So, positive/negative rating drift values represent situations where the user's submitted rating was higher/lower than the system's predicted rating. The left side of Table 8 shows the mean values across the four conditions of the study.

We began with the general first-order linear regression model, paralleling the preliminary analyses for study 1. In study 2, the rating drift of the viewed episode is the dependent variable. There is a single manipulated anchoring factor with three experimental levels (i.e., high, low, and accurate). Age and the ratings of the five genres are included as covariates; also, the hours of television watched and gender are added as controls, as in study 1. To account for potential effects resulting from participants watching different television shows, we included several TV series and episode-level controls in the analysis.

Table 8 **Mean (Standard Deviation) Rating Drift of the Viewed TV Show by Experimental Condition, Overall and for Subjects Who Watched the Same Show, Study 2**

	Overall		Single show	
Group	*N*	Mean (SD)	*N*	Mean (SD)
High	51	0.40 (1.00)	27	0.81 (0.82)
Control	48	0.14 (0.94)	22	0.53 (0.76)
Accurate	51	0.13 (0.96)	27	0.37 (0.93)
Low	47	−0.12 (0.94)	26	0.30 (0.86)

Specifically, we estimated three alternative regression models to demonstrate the robustness of the results. Models 1 and 2 perform analyses on the full data set, including the treatment levels and basic participant-level controls discussed above. For both, we employ robust standard errors clustered on the TV series watched to account for any corresponding error correlations. Model 1 also includes control variables on the general appeal of the TV show watched, captured from Hulu.com at the time of the experiment, to account for show differences. These include the average rating for the television series (not specific episode), the number of ratings the series received on Hulu.com, the number of episodes of the TV show available for viewing on Hulu.com, and the number of views the TV series received on Hulu.com. As a robustness check, model 2 mimics model 1, except the show-level control variables from Hulu.com were replaced with data captured from the Internet Movie Database (IMDB.com) on the specific TV episode watched. In model 2, two new control variables are introduced: the IMDB average rating for the specific episode watched and the number of ratings the specific episode had received on IMDB. IMDB data were captured in September 2012. Model 3 utilizes the standard OLS regression, but applied to a subsample ($n = 102$) of the data in which all participants watched the same TV episode, thereby controlling for episode factors in a different manner. The full results of all models are presented in Table 9.

Anchoring treatment effects were consistently significant across the three models. Of the secondary variables in models 1 and 2, the show and episode controls from Hulu.com and IMDB had significant effects; however, the magnitudes of the effects were extremely small. None of the other secondary effects were consistent across all three models. For the treatment, the anchor manipulations had a statistically significant impact ($p < 0.01$) in all three models. Participants who saw a high perturbation in the system rating had positive rating drift of about one half-star higher, on average, as compared to participants who received low perturbations (e.g., model 1 coefficient = 0.484). We turn to the planned contrasts to more directly test the Anchoring Hypothesis, as well as to test for any asymmetry of the effect.

Figure 4 summarizes the direct contrasts of interest. Providing an accurate recommendation did not significantly affect preferences for the show, as compared to the control condition (two-tailed $t(97) = 0.023$, $p = 0.982$). In other words, providing nonperturbed recommendations does not produce anchoring effects. Although this is not necessarily surprising and does not eliminate the possibility that anchoring effects can still occur on an *individual* item basis (e.g., for types of items where the recommender system may have

Table 9 Analysis of Study 2 Data

		Dependent variable: Rating drift		
		Model 1 Full, Hulu	*Model* 2 Full, IMDB	*Model* 3 Single show
Treatments		Coefficient (Std. error)	Coefficient (Std. error)	Coefficient (Std. error)
Anchoring (*Baseline:Low*)				
High		0.484 (0.144)**	0.530 (0.129)***	0.632 (0.234)**
Accurate		0.184 (0.127)	0.261 (0.134)^	0.209 (0.244)
Constant		−2.239 (1.783)	−2.696 (1.219)*	−1.572 (0.911)^
Control variables				
Age		0.0145 (0.007)^	0.017 (0.003)***	0.020 (0.014)
Is Male		0.282 (0.175)	0.342 (0.153)*	0.362 (0.212)^
Hours Watched	> 20	−1.692 (0.956)	−1.935 (0.742)*	−3.026 (1.022)**
(Baseline: < 1)	12–16	−0.035 (0.362)	−0.045 (0.365)	−0.305 (0.384)
	4–8	−0.191 (0.297)	−0.241 (0.249)	−0.622 (0.276)*
	2–4	−0.200 (0.271)	−0.211 (0.262)	−0.537 (0.267)*
Comedy		0.029 (0.084)	0.076 (0.093)	0.076 (0.130)
Drama		0.162 (0.115)	0.173 (0.119)	0.188 (0.141)
Mystery		0.091 (0.064)	0.083 (0.068)	0.135 (0.132)
Reality		−0.007 (0.046)	0.000 (0.050)	−0.032 (0.092)
Sci-Fi/Fantasy		0.159 (0.074)^	0.157 (0.067)*	0.100 (0.091)
Hulu Avg Show Rating		0.243 (0.238)		
Hulu # of Ratings (Show)		−0.001 (0.000)*		
Hulu Number of Episodes		0.007 (0.003)*		
Hulu Number of Views		−0.014 (0.005)*		
IMDB Avg Episode Rating			0.016 (0.191)	
IMDB # of Ratings (Episode)			0.001 (0.000)***	
R^2		0.330	0.316	0.346

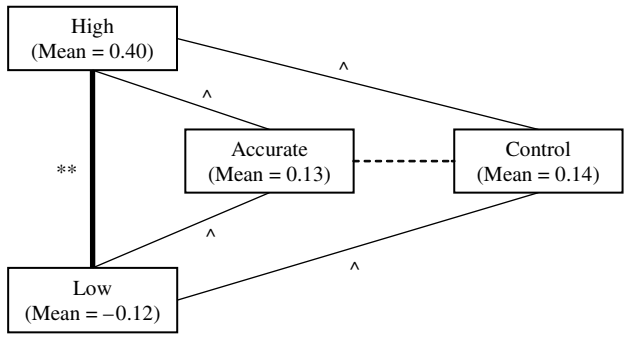
Notes. Model Descriptions: (1) Includes all data from Study 1, participant-level controls, clustered errors based on show, and TV show control variables based on Hulu.com ratings; (2) includes all observations, participant-level controls, clustered standard errors based on show, and TV episode control variables passed on IMDB (imdb.com) ratings; (3) includes participant-level controls and only observations for participants who viewed the same TV show.

Significance levels: ^ ≤ 0.1, * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

a systematic overprediction or under-prediction bias due to lack of data), it is nevertheless important to know that "best effort" recommendations (i.e., unbiased, nonperturbed recommendations generated by a state-of-the-art algorithm) *on average* do not lead to anchoring effects. As with study 1, the high recommendation condition led to inflated ratings compared to the low condition (one-tailed $t(96) = 2.629$, $p = 0.005$). 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 ends. There was a marginally significant effect of the recommendation being lowered compared to being accurate ($t(96) = 1.305$, $p = 0.098$, Cohen's $d = 0.30$), and a marginally significant effect at the high end compared to receiving accurate recommendations ($t(100) = 1.366$, $p = 0.088$, Cohen's $d = 0.23$). Similarly, providing a recommendation that was raised or lowered, compared to not providing any recommendation, led to increased preference ($t(97) = 1.391$, $p = 0.084$) and decreased preference ($t(93) = 1.382$, $p = 0.085$), respectively. In summary, the Anchoring Hypothesis is supported in study 2 consistently with study 1; however, the anchoring effects were observed to be 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 participants. We included control variables to account for differences in popularity of the shows watched (Table 9); however, additional idiosyncrasies of each show may still impact the results. As noted, 102 of the 198 participants in study 2 (52%) watched the same TV show episode and, as a result, we were able to perform post-hoc analyses, paralleling the main analyses, limited to this subset of viewers. The mean values across the four conditions for the main response variable are shown in Table 8, and the results of the general linear regression model (applied to the restricted sample) are provided as model 3 in Table 9. Of main interest is that the primary anchoring manipulation continued to have a statistically significant effect when comparing high versus low perturbations

Figure 4 Contrasts of Experimental Conditions in Study 2



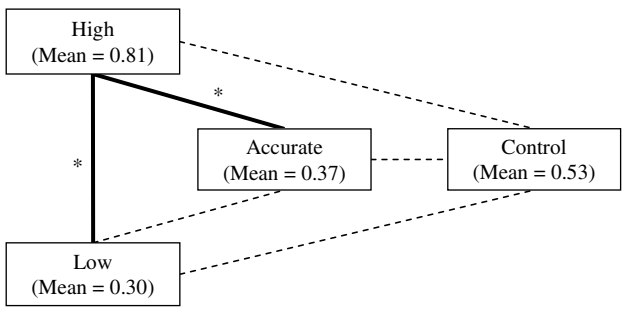
$^{\wedge}p < 0.1$, $^{**}p < 0.01$.

($p = 0.010$). We turn to the analysis of the planned contrasts to identify the nature of the effects.

Figure 5 summarizes the corresponding direct contrast results for the restricted data set (paralleling Figure 4 for the full data set). Providing an accurate recommendation still did not significantly affect preferences for the show, as compared to the control condition (two-tailed $t(47) = 0.671$, $p = 0.506$). Consistent with study 1 and the overall analyses, the high recommendation condition led to inflated ratings compared to the low condition (one-tailed $t(51) = 2.213$, $p = 0.016$). 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 ends were not symmetric. Although the control (i.e., not providing a recommendation) showed no significant difference from either a high ($t(47) = 1.210$, $p = 0.116$) or low recommendation ($t(46) = 0.997$, $p = 0.162$), compared to receiving an accurate recommendation there was a significant effect of the recommendation being raised ($t(52) = 1.847$, $p = 0.035$, Cohen's $d = 0.50$), but not of being lowered ($t(51) = 0.286$, $p = 0.388$).

Thus, the indicated asymmetry of the anchoring effect is different from the asymmetry present in study 1, presenting at the high end rather than the low end. Also, the asymmetry is not robust across the overall data, indicating that the underlying cause of asymmetries is situational, in this case depending upon specific TV show effects. Our study was exploratory on this issue and, clearly, a separate, dedicated study is needed to differentiate the possible effects. To gain more data and further explore the granularity of the anchoring effect, we included the use of both high and low anchors in the design of study 3.

Figure 5 Contrasts of Experimental Conditions for Subjects Watching the Same TV Show in Study 2



$^{*}p < 0.05$.

5. Study 3: Fine-Tuning the Effects of Real Recommendations

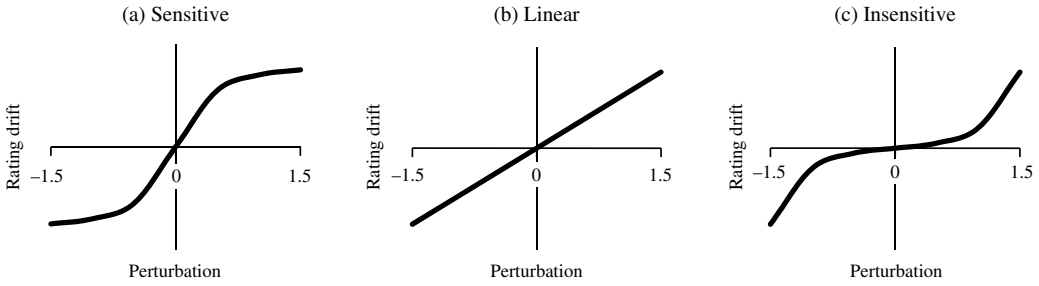
Study 3 provides several contributions that extend the findings of studies 1 and 2. The primary goal was to investigate the *granularity* of the observed anchoring effects. Study 2 showed the impact of large perturbations of recommendations provided by a commonly used algorithm, but did not inform us of the functional form of the effect. To demonstrate the issue, consider the three possibilities illustrated by Figure 6. All three functional forms contain the same three points at perturbation levels of −1.5, 0, and 1.5, with rating drift values consistent with the results of study 2. However, the specific patterns differ among the three displayed functions, any of which is plausible. Figure 6(a) shows a situation in which the respondent is highly sensitive to perturbations, exhibiting strong effects even for small changes. Figure 6(b) shows a respondent who is consistently sensitive, exhibiting a linear pattern. Figure 6 (c) is of an insensitive respondent, who only shows an effect for larger changes and is relatively unaffected by smaller changes.

Therefore, the primary contribution of study 3 is to better understand the function underlying the broad-level effects observed in study 2 by providing a finer analysis using a continuum of perturbations. Secondly, we employ a within-subjects design in study 3, allowing us to better investigate and control for individual differences. Finally, recall that our interest in this research is to test effects of anchoring at the time of consumption. Thus, to explore the functional relationships and the within-subjects effects, we needed to use stimuli other than television shows that allowed for multiple consumption events for each participant in a reasonable amount of time. So, additionally, a third benefit of study 3 is that we could test the Anchoring Hypothesis in a different context (using jokes as stimuli), providing evidence of the generalizability of our results to a different content domain.

5.1. Methods

5.1.1. Participants. Sixty-two people completed the study. They were solicited from the same paid

Figure 6 **Three Contrasting Functional Forms Consistent with Anchoring Effects Observed in Studies 1 and 2**



participant pool as used for studies 1 and 2 with no overlap across the three studies. Participants received a fixed fee upon completion of the study. One person rated the jokes uniformly low, so we were unable to provide the full range of manipulations for the study. This subject was removed from the analyses for consistency, leaving 61 subjects. Demographic features of the sample are summarized in the third data column of Table 3. The samples are comparable across the three studies.

5.1.2. Procedure. As with study 2, the anchors received by subjects were based on the recommendations of a true real-time recommender system. The item-based CF technique was used to maintain consistency with study 2 and because the technique tends to perform well (Adomavicius and Tuzhilib 2005, Deshpande and Karypis 2004, Sarwar et al. 2001). A list of 100 jokes was used for the study, with the order of the jokes randomized across participants. The jokes and the rating data for training the recommendation algorithm were taken from the Jester Online Joke Recommender System repository, a database of jokes and preference data maintained by the University of California, Berkeley (http://eigentaste.berkeley.edu/dataset). Specifically, we used their data set 2 of 150 jokes. To get to our list of 100, we removed the 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; see Goldberg et al. 2001), 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. Participants first evaluated 50 jokes, randomly selected from the list of 100 and randomly ordered, to form a basis for providing recommendations. The same five-star rating scale from studies 1 and 2 was used, allowing half-star ratings. Next, the subjects received 40 jokes with a predicted system rating displayed next to the input drop-down box for providing a preference rating. Thirty of these predicted ratings were perturbed, five 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 five-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 jokes, 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 (three sessions were used in total), 10 more jokes were added as a control in random order (i.e., with no predicted system rating provided). Finally, in all sessions participants completed a short demographic survey.

5.2. Results

As with study 2, the main response variable for study 3 was rating drift, which was defined in the same manner (i.e., actual rating–predicted rating). Before turning to the primary analysis, we check that the aggregate data are consistent with those of study 2. For comparison purposes, Table 10 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 of results for study 3—using jokes and a within-subjects design—parallels that for study 2 (compare to Table 8)—using TV shows and a between-subjects design.

As an initial step in our analysis, we created a panel from the data. The repeated measures design of study 3, wherein each participant was exposed to all treatment levels in a random fashion, allows us to model the aggregate relationship between rating perturbations and rating drift while controlling for individual participant differences. The standard OLS model using robust standard errors, clustered by participant, and participant-level controls represents our

Table 10 Mean (SD) Rating Drift of the Jokes in Study 3, Under Comparable Conditions used in Study 2 (±1.5, 0, Control)

	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)

baseline model for the analysis of study 3 (see model 1 in Table 11):

$$\begin{aligned}&\text{RatingDrift}_{ij}\\&=b_0+b_1(\text{Perturbation}_{ij})+b_2(\text{PredictedRating}_{ij})\\&\quad+b_3(\{\text{ParticipantControls}\}_i)+b_4(\text{AverageJokeRating}_j)\\&\quad+b_5(\text{AverageJokeRating}_j\times\text{Perturbation}_{ij})+\varepsilon_{ij}.\end{aligned}$$

The data panel consists of $i=1,\ldots,n$ participants with $j=1,\ldots,m$ observations per participant, in a balanced panel. The main effect of interest is the relationship between the rating perturbation (treatment) and the resulting rating drift (dependent variable). Unobserved differences across jokes may also impact rating drift; we therefore control for the predicted rating for each joke as well as any joke effects in our model. We also control for the average joke rating in the Jester database and any interaction this could have with the introduced perturbation.

The results demonstrate a significant positive relationship between perturbation and rating drift, while controlling for individual participant factors, joke effects, and the predicted rating (Table 11). The coefficient for perturbation (0.3475, $p < 0.001$) provides strong evidence that the participants' rating drift correlates with the perturbations in recommendations. Of the secondary factors, only gender and average joke ratings had significant effects. Therefore, controlling for the joke effect in the regression model was warranted and the significant aggregate main effect of perturbation on rating drift can be considered robust.

We estimated two additional regression models to account for potential sources of heterogeneity in the data. First, a random effects model was tested (not shown), in which we incorporated observation-invariant variables from the post-experiment survey (e.g., age, gender, etc.). The results of the random effects analysis were identical in terms of the relationship between perturbation and rating drift; however, the Hausman test for the appropriateness of a random effects model showed strong correlation between the regressors and the individual participant effects, indicating that the fixed-effects model is preferred. Consequently, we estimated a fixed-effects model (model 2 in Table 11) that included an individual participant fixed effect u_i in lieu of the participant controls. We note here that we also tested a model that included the number of ratings for each joke as a control variable; however, the number of ratings was highly correlated with the average rating ($r = 0.98$) and was ultimately omitted for multicollinearity reasons. These additional models account for participant-level heterogeneity (with either fixed effects or participant-level controls) and joke-level heterogeneity (with joke-level controls). The results of model 2 also show that there were significant individual participant effects ($F(60, 2{,}375) = 2.77$, Prob > $F = 0.0000$), and these effects account for about 15.5% ($\rho = 0.1552$) of the variation observed in the rating drift. In all models, the effect of the perturbation on rating drift was highly significant (at the 0.001 level) and had approximately the same magnitude, which provides strong evidence of the robustness of the results.

Table 11 Study 3 Initial Regression Analysis

	Model 1 CE	*Model* 2 FE
Predictor variables	Coefficient (Std. error)	Coefficient (Std. error)
Perturbation	0.3475 (0.0324)***	0.3478 (0.0187)***
Predicted Rating	−0.0227 (0.0877)	0.8234 (0.3923)*
Constant	1.3118 (0.3402)	−1.6087 (0.8729)^
Control variables		
Age	0.0033 (0.0040)	
Recs. Accurate	−0.0282 (0.0371)	
Recs. Useful	0.0291 (0.0248)	
Is Male	0.1198 (0.0540)*	
Is Grad Student	−0.0335 (0.0767)	
Is Native Speaker	0.0039 (0.0691)	
Prior Rec Sys Exper.	0.0293 (0.0563)	
Average Joke Rating	−0.4668 (0.1077)***	−1.2880 (0.3818)**
Avg Joke Rating × Perturbation Interaction	0.0609 (0.0608)	0.0651 (0.0587)
Participant fixed effects		
σ_u^2		0.3703
σ_ε^2		0.8640
ρ		0.1552
F-Statistic (test that all $u_i = 0$)		2.77***
Model fit		
R^2 *Overall*	0.1736	0.1108
R^2 *Within*		0.1781
R^2 *Between*		0.0002
F-Statistic	27.38***	129.17***

Notes. Model Descriptions: (1) Includes robust standard errors clustered by participant (CE) as well as joke and participant-level control variables; (2) includes participant-level fixed effects (FE), as well as joke and participant-level control variables.

$N = 2{,}440$ observations, $n = 61$ participants, $m = 40$ observations per participant.

Significance Levels: ^ ≤ 0.1, * ≤ 0.05, ** ≤ 0.01, *** ≤ 0.001.

Figure 7 **Aggregated Mean Rating Drift as a Function of the Amount of Perturbation of the System Recommendation** (−1.5, −1, −0.5, 0, 0.5, 1, 1.5) **and for the Control Condition in Study 3**

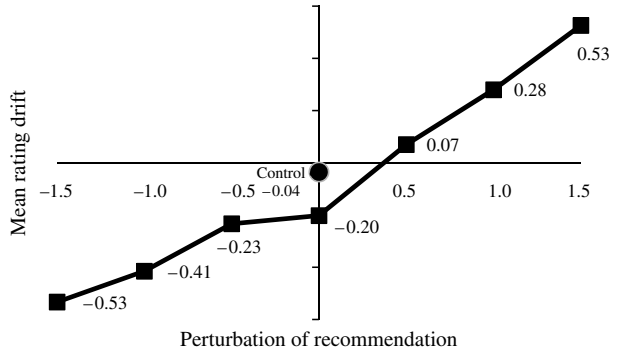


Figure 7 shows the mean rating drift, aggregated across items and subjects, for each perturbation used in the study. In aggregate, there is a linear relationship both for negative and positive perturbations. However, this does not necessarily imply that all the individual respondents share this functional form. We investigate individual responses in the next analyses, taking advantage of the within-subjects design.

The detailed analysis applies regression modeling for the data from each individual. For every participant, there are five joke observations for each of the six perturbation sizes (−1.5, −1.0, −0.5, +1.5, +1.0, and +0.5) and 10 jokes whose recommendations were not perturbed (perturbation = 0). Thus, for each respondent, there are $n = 40$ observations (jokes) potentially available for analysis. Since studies 1 and 2 indicated a possible asymmetry of the effects for positive and negative perturbations, we begin by checking the functional form for each individual separately for positive and negative perturbations.

First, to distinguish among the three possible functional forms illustrated in Figure 6, we performed two regression analyses for each participant, one on the positive side ($n = 25$, including responses with perturbations of 0, +0.5, +1, and +1.5), and one on the negative side ($n = 25$, including responses with perturbations of 0, −0.5, −1, and −1.5). Since the regression analyses were done separately and without any intention of directly comparing the two sets of analyses, the 10 common data points with no perturbations were included in both the negative and positive analyses. This allowed for maximal sensitivity from the available data for any curvature that may be present in the functional relationships.

On the positive side for each of the 61 respondents, a regression model was fit ($n = 25$) using rating drift as the dependent variable and perturbation size as the independent variable in a first-order, quadratic model:

$$\text{Rating Drift} = b_0 + b_1(\text{Perturbation}) + b_2(\text{Perturbation})^2 + \varepsilon.$$

Note that, since these are individual level regressions, no additional covariates are available to be included in the model. The test of $b_2 = 0$ (the quadratic term) offers evidence of curvature in the relationship. A similar regression was performed separately on the negative side ($n = 25$) for each of the 61 respondents. On the positive side, of the 61 subjects, only four demonstrated a significant curvilinear relationship even using a liberal criterion of $\alpha = 0.10$, two showing the pattern in Figure 6(a) and two like that in Figure 6(c). Since we would expect about 6.1 respondents to exhibit a significant effect at random using $\alpha = 0.10$, no significant pattern of curvilinearity across the respondents can be concluded. On the negative side, of the 61 subjects, only eight demonstrated a significant curvilinear relationship using the liberal criterion of $\alpha = 0.10$ (of these, only four were significant at $\alpha = 0.05$). The types of curvature were evenly split for the eight (and the four) between the patterns of Figures 6(a) and 6(c). So, again, no significant pattern of curvilinearity across respondents was concluded. The linear relationship was accepted as the predominant pattern.

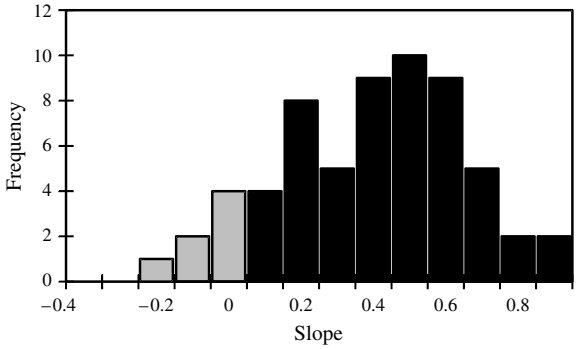
The second step of the individual-level analysis was to test for asymmetry (i.e., differences in the linear slopes) between the positive and negative perturbations. Since this involves a direct comparison between the positive and negative responses, only the perturbed values were used in this analysis. In each direction, positive and negative, there are five observations for each perturbation size—1.5, 1.0, and 0.5—so the sample size for each regression is $n = 15$. Each of the 61 participants was analyzed using two first-order linear regression analyses (positive and negative), producing two slopes (b_1) for each subject:

$$\text{Rating Drift} = b_0 + b_1(\text{Perturbation}) + \varepsilon.$$

The mean positive slope (0.46) was slightly higher than the mean negative slope (0.30). Using a paired difference test, no significant difference was observed between the slopes for positive and negative perturbations ($t(60) = 1.39$, two-tailed $p = 0.17$). Thus, the most likely functional form of the relationship between perturbation size and rating drift is illustrated by Figure 6(b). Our analyses suggest that the relationship is linear with a similar slope across the positive and negative instances of perturbation size.

Consequently, for the last step of the analyses, we combine the data to obtain a single fit for the simple linear regression model from each subject. For the combined analysis, the nonperturbed data (10 jokes) were returned, bringing the sample size to $n = 40$ for each individual's regression. Figure 8 shows the distribution of slope values across participants. The mean slope value across the 61 participants is 0.35,

Figure 8 **Distribution of Slopes by Subject, for Analysis of Rating Drift as a Linear Function of Rating Perturbations ($-1.5, -1, -0.5, 0, +0.5, +1,$ and $+1.5$)**



Notes. All regressions based on $n = 40$. Zero and negative slope values are in gray.

and is significantly positive ($t(60) = 10.74$, two-tailed $p < 0.0001$). At an individual level, only one participant had a significantly negative slope at a two-tailed cutoff level of 0.10. Five other participants had a nonsignificantly negative slope, and one subject had a 0 slope. In contrast, 43 participants had a significantly positive slope, with an additional 11 participants having a nonsignificantly positive slope. Thus, the positive effect between the perturbations and the rating drift is observed for most individuals. The effect is linear and symmetric across the full range of perturbations used.

6. Discussion and Conclusions

6.1. Summary of Contributions

In this study, we conducted three laboratory experiments and systematically examined the impact of recommendations on consumer preferences. The study integrates ideas from behavioral decision theory and recommender systems, both from practical and theoretical standpoints. Experimental results provide strong evidence that biased output from recommender systems can significantly influence consumers' preference ratings. Using both artificial and real recommendations, and two different content domains (TV shows and jokes), the primary Anchoring Hypothesis was consistently confirmed. Users' preference ratings are malleable and sensitive to the recommendations received; the expressed preference is pulled toward the value of the provided recommendation.

In terms of the Anchoring Hypothesis, the studies extend the existing, yet limited, research in the domain of preference construction. Further, we demonstrate that the anchoring effect operates even at the point of consumption. Thus, anchoring not only impacts recall tasks where uncertainty is high and biases are expected (e.g., Cosley et al. 2003), but also impacts ratings at the point of integrating information and formulating a response, where uncertainty and biased-recall explanations are not operable. In addition, the effect is prompted not only at discrete, extreme manipulations; it was observed to be continuous, with the magnitude of the drift in ratings proportional to the magnitude of the perturbation of the recommendation.

The Timing Hypothesis, and thus the priming explanation, was not supported. In contrast, the System Reliability Hypothesis was supported, consistent with an explanation that the recommendation is viewed as a suggestion to a "correct" answer. When strength of the recommendation was reduced by stating that the recommender system was tentative and in early stages of development, the recommendation had a lesser effect than when provided by a system without such qualifications. The effect is not simply numerically driven; the recommendation's reliability has an impact. This supports the direction of recent research (Komiak and Benbasat 2006, Wang and Benbasat 2007), tying the use of recommender agents to trust attitudes. The value of further research along these lines is strongly indicated.

Finally, the studies indicate that the impact of perturbing the ratings is not necessarily symmetric (i.e., when the recommendation is adjusted upward versus downward); and the asymmetry is situationally dependent. We observed a symmetric pattern of anchoring effects when analyzing *aggregate* impact across multiple items, i.e., across a variety of TV shows in study 2 (aggregate analysis, Figure 4) and across a variety of jokes in study 3 (Figure 7). In contrast, we observed different asymmetric patterns when analyzing anchoring impact on *specific* items, i.e., for a specific TV show in both study 1 (Figure 2) and study 2 (single show analysis, Figure 5). This difference may tie to situational effects on trust or to other mechanisms. The pattern provides a promising basis for additional research along these lines.

6.2. Practical Implications and Future Work

As shown by Figure 1, recommender systems are designed to incorporate a feedback loop between user-reported ratings and system-generated recommendations. Anchoring effects potentially impact this feedback loop and, thus, the use and design of recommender systems. From a practical perspective, the findings of our research have several important implications. First, standard performance metrics for recommender systems may need to be adjusted to account for these anchoring-related phenomena. If recommendations can influence consumer-reported ratings, then how should recommender systems be objectively evaluated? The system reliability finding

suggests that consumers have some inherent trust in recommender systems that may or may not be warranted. This result raises the possible opportunity for third party organizations to rate and ensure the efficacy of recommender systems as means of reinforcing consumer trust, not unlike the use of Better Business Bureau seals to signal quality to consumers. Second, the results suggest the potential of anchoring effects biasing the inputs to recommender systems. If two consumers give the same rating based on different initial recommendations, do their preferences really match in identifying future recommendations? Third, our findings bring to light the potential impact of recommendations on strategic 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 (Shih et al. 2007). If consumer choices are significantly influenced by recommendations, regardless of accuracy, then there is a potential for firms to leverage recommender systems for competitive advantage.

In the areas of recommender system algorithm and user interface design, there is significant potential for developers to create recommender system designs that compensate for anchoring effects, thereby improving recommendation performance and use. In particular, since the user's preference rating for an item can be manipulated just by observing a recommendation before (or after) item consumption, a "smart" interface design would determine whether the user has encountered a recommendation for the item before submitting his or her preference rating. A similar problem arises in online advertising when using a cost-per-action (CPA) pricing model, where the advertiser pays for the ad only when the ad leads to a desired conversion action or event, e.g., a purchase or subscription (Wilson and Pettijohn 2010). One of the key elements of this model is the use of "conversion tracking" that allows advertisers to reliably track the conversion action back to the initial click on an ad. The integration of similar techniques with recommender system interfaces, in order to determine whether user's preference rating submission was preceded by an observation of a recommendation, represents a promising direction for future work. Another interesting research direction would be to explore whether various alternative interface designs (Yoo and Gretzel 2011, Tintarev and Masthoff 2011), including ones that provide recommendations without presenting numeric ratings to users, may reduce (or eliminate) anchoring effects.

Also, as pointed out in the paper, the "best effort" recommendations (i.e., unbiased, nonperturbed recommendations) on average do not produce anchoring effects (relative to a control). Therefore, in terms of algorithm design, determining whether the observed recommendation was biased constitutes another possible direction for future work. System-generated ratings can be biased for two reasons: (i) because of inherent limitations of the input data or predictive modeling technique that consistently overpredicts or underpredicts ratings for certain types of items, and (ii) because of manipulations from malicious users (Lam and Riedl 2004, Mobasher et al. 2007). Better understanding of systematic bias in recommendation techniques and the application of attack-detection approaches could prove to be valuable ingredients for future improvements to recommender systems, achieved by compensating for anchoring effects. Of course, before redesigning recommender systems' algorithms or interfaces, one needs to (a) demonstrate that such anchoring effects exist and (b) measure their magnitude. This was exactly the focus of our studies.

In conclusion, further research is clearly needed to understand the effects of recommender systems on consumer preferences and behavior. Issues of trust, bias, and preference realization appear to be intricately linked in the use of recommendations in online marketplaces. Additionally, the situational asymmetry of these effects must be explored to understand what characteristics have the largest influence. Moreover, future research is needed to investigate the error compounding issue: If a recommender system keeps providing biased recommendations, how far can people be pulled? Our studies have 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 from both design science and social science perspectives.

Acknowledgments

The authors thank the senior editor, the associate editor, and the anonymous review team for their thoughtful guidance during the review process. This work was supported in part by the National Science Foundation [Grant IIS-0546443].

References

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

Ariely D, Lewenstein G, Prelec D (2003) "Coherent arbitrariness": Stable demand curves without stable preferences. *Quart. J. Econom.* 118:73–105.

Bell RM, Koren Y (2007) Improved Neighborhood-based Collaborative Filtering. Berkhin P, Caruana R, Wu X, eds. *KDD Cup and Workshop* (ACM, New York), 7–14.

Bennet J, Lanning S (2007) The Netflix Prize. Berkhin P, Caruana R, Wu X, eds. *KDD Cup and Workshop* (ACM, New York), 3–6.

Breese JS, Heckerman D, Kadie C (1998) Empirical analysis of predictive algorithms for collaborative filtering. Cooper G, Moral S, eds. *Fourteenth Conf. Uncertainty in Artificial Intelligence, Madison, WI* (Morgan Kaufmann, San Francisco), 43–52.

Chapman G, Bornstein B (1996) The more you ask for, the more you get: Anchoring in personal injury verdicts. *Appl. Cognitive Psych.* 10:519–540.

Chapman G, Johnson E (2002) Incorporating the irrelevant: Anchors in judgments of belief and value. Gilovich T, Griffin D, Kahneman D, eds. *Heuristics and Biases: The Psychology of Intuitive Judgment* (Cambridge University Press, Cambridge, UK), 120–138.

Cohen J (1988) *Statistical Power Analysis for the Behavioral Sciences*, 2nd ed. (Lawrence Erlbaum Associates, Hillsdale, NJ), 273–288.

Cosley D, Lam S, Albert I, Konstan JA, Riedl J (2003) Is seeing believing? How recommender interfaces affect users' opinions. Cockton G, Korhonen P, eds., *CHI 2003 Conference, Fort Lauderdale FL* (ACM, New York), 585–592.

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

Epley N, Gilovich T (2010) Anchoring unbound. *J. Consumer Psych.* 20:20–24.

Flynn LJ (2006) Like this? You'll hate that. (Not all Web recommendations are welcome.) *New York Times*, (January 23) http://www.nytimes.com/2006/01/23/technology/23recommend.html.

Funk S (2006) *Netflix Update: Try This at Home*, (December 11, 2006). Retrieved January, 2010), http://sifter.org/~simon/journal/20061211.html.

Goldberg K, Roeder T, Gupta D, Perkins C (2001) Eigentaste: A constant time collaborative filtering algorithm. *Inform. Retrieval* 4(July):133–151.

Jacowitz KE, Kahneman D (1995) Measures of anchoring in estimation tasks. *Personality Soc. Psych. Bull.* 21:1161–1166.

Johnson JEV, Schnytzer A, Liu S (2009) To what extent do investors in a financial market anchor their judgments excessively? Evidence from the Hong Kong horserace betting market. *J. Behav. Decision Making* 22:410–434.

Komiak S, Benbasat I (2006) The effects of personalization and familiarity on trust and adoption of recommendation agents. *MIS Quart.* 30(4):941–960.

Koren Y (2009) The BellKor solution to the Netflix grand prize. *NetflixPrize.com*. http://www.netflixprize.com/assets/GrandPrize2009_BPC_BellKor.pdf.

Koren Y, Bell R, Volinsky C (2009) Matrix factorization techniques for recommender systems. *IEEE Comput. Soc.* 42:30–37.

Ku G, Galinsky AD, Murnighan JK (2006) Starting low but ending high: A reversal of the anchoring effect in auctions. *J. Personality Soc. Psych.* 90:975–986.

Lam S, Riedl J (2004) Shilling recommender systems for fun and profit. Feldman S, Uretsky M, eds. *13th Internat. Conf. World Wide Web, New York City, NY* (ACM, New York), 353–402.

Lichtenstein S, Slovic P, eds. (2006) *The Construction of Preference* (Cambridge University Press, Cambridge, UK).

McNee SM, Lam SK, Konstan JA, Riedl J (2003) Interfaces for eliciting new user preferences in recommender systems. *User Modeling 2003, Proc.*, (Springer-Verlag, Berlin), 178–187.

Mobasher B, Burke R, Bhaumik R, Williams C (2007) Toward trustworthy recommender systems: An analysis of attack models and algorithm robustness. *ACM Trans. Internet Tech.* 7(4):Article 23.

Mussweiler T, Strack F (2000) Numeric judgments under uncertainty: The role of knowledge in anchoring. *J. Experiment. Soc. Psych.* 36:495–518.

Northcraft G, Neale M (1987) Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organ. Behav. Human Decision Processes* 39:84–97.

Piotte M, Chabbert M (2009) The pragmatic theory solution to the Netflix grand prize, (August, 2009), http://www.netflixprize.com/assets/GrandPrize2009_BPC_PragmaticTheory.pdf.

Pu P, Chen L (2007) Trust-inspiring explanation interfaces for recommender systems. *Knowledge-Based Systems* 20(August 6): 542–556.

Russo JE (2010) Understanding the effect of a numerical anchor. *J. Consumer Psych.* 20:25–27.

Sarwar B, Karypis G, Konstan JA, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. Shen V, Saito N, Lyu M, Zurko M, eds., *10th Intl. WWW Conf., Hong Kong* (ACM, New York), 285–295.

Schkade DA, Johnson EJ (1989) Cognitive processes in preference reversals. *Organ. Behav. Human Decision Processes* 44:203–231.

Schonfeld E (2007) Click here for the upsell. *CNNMoney.com* (July). http://money.cnn.com/magazines/business2/business2_archive/2007/07/01/100117056/index.htm.

Shih W, Kaufman S, Spinola D (2007) Netflix. *Harvard Bus. School*, Case 9-607-138, May 31.

Strack F, Mussweiler T (1997) Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. *J. Personality Soc. Psych.* 73:437–446.

Swearingen K, Sinha R (2001) Beyond algorithms: An HCI perspective on recommender systems. Herlocker J, ed., *ACM SIGIR 2001 Workshop on Recommender Systems, New Orleans, LA*.

Thorsteinson T, Breier J, Atwell A, Hamilton C, Privette M (2008) Anchoring effects on performance judgments. *Organ. Behav. Human Decision Processes* 107:29–40.

Tintarev N, Masthoff J (2011) Designing and evaluating explanations for recommender systems. Ricci F, Rokach L, Shapira B, Kantor P, eds. *Recommender Systems Handbook* (Springer, New York), 479–510.

Töscher A, Jahrer M, Bell R (2009) The bigchaos solution to the Netflix grand prize, (September 5, 2009), http://www.netflixprize.com/assets/GrandPrize2009_BPC_BigChaos.pdf.

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

Wang W, Benbasat I (2007) Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *J. Management Inform. Systems* 23(4):217–246.

Wilson R, Pettijohn J (2010) Tracking online ad campaigns: A primer. *J. Direct, Data and Digital Marketing Practice* 12:69–82.

Yoo K, Gretzel U (2011) Creating more credible and persuasive recommender systems: The influence of source characteristics on recommender system evaluations. Ricci F, Rokach L, Shapira B, Kantor P, eds. *Recommender Systems Handbook* (Springer, New York), 455–477.