REDUCING RECOMMENDER SYSTEM BIASES: AN INVESTIGATION OF RATING DISPLAY DESIGNS[1]

Gediminas Adomavicius
Information and Decision Sciences, Carlson School of Management, University of Minnesota,
321 19th Avenue South, Minneapolis, MN 55455 U.S.A. {gedas@umn.edu}

Jesse C. Bockstedt
Information Systems and Operations Management, Goizueta Business School, Emory University,
1300 Clifton Road, Atlanta, GA 30322 U.S.A. {bockstedt@emory.edu}

Shawn P. Curley
Information and Decision Sciences, Carlson School of Management, University of Minnesota,
321 19th Avenue South, Minneapolis, MN 55455 U.S.A. {curley@umn.edu}

Jingjing Zhang
Operations and Decision Technologies, Kelley School of Business, Indiana University,
1309 East Tenth Street, Bloomington, IN 47405 U.S.A. {jjzhang@indiana.edu}

Prior research has shown that online recommendations have a significant influence on consumers' preference ratings and economic behavior. Specifically, biases induced by observing personalized system recommendations can lead to distortions in users' self-reported preference ratings after consumption of an item, thus contaminating the users' subsequent inputs to the recommender system. This, in turn, provides the system with an inaccurate view of user preferences and opens up possibilities of rating manipulation. As recommender systems continue to become increasingly popular in today's online environments, preventing or reducing such system-induced biases constitutes a highly important and practical research problem. In this paper, we address this problem via the analysis of different rating display designs for the purpose of proactively preventing biases before they occur (i.e., at rating collection time). We use randomized laboratory experimentation to test how the presentation format of personalized recommendations affects the biases generated in post-consumption preference ratings. We demonstrate that graphical rating display designs of recommender systems are more advantageous than numerical designs in reducing the biases, although none are able to remove biases completely. We also show that scale compatibility is a contributing mechanism operating to create these biases, although not the only one. Together, the results have practical implications for the design and implementation of recommender systems as well as theoretical implications for the study of recommendation biases.

Keywords: Recommender systems, decision bias, interface design, preference ratings, scale compatibility, experimental research

---

Introduction ▪▪▪▪▪▪▪

Recommender systems are increasingly used to curate content for individuals in a broad variety of application domains, including e-commerce, media, business, education, tourism, and government (Lu et al. 2015). They help consumers effectively find relevant information, products, and services among a potentially huge number of choices. They are also beneficial to companies (i.e., providers); by providing value to consumers, effective recommender systems can serve as a differentiator from competitors as well as increase consumer loyalty and consumption levels. For example, it is estimated that 30% of Amazon's page views result from recommendations (Sharma et al. 2015); similarly, Netflix has reported that over 80% of the content watched by its subscribers was suggested by its recommendation systems (Gomez-Uribe and Hunt 2016). Given the ubiquity of recommender systems and their increasing role in consumer decision making, understanding the effects and limitations of such information systems is of immense practical importance.

A key component of recommender systems' success is their ability to collect information on user preferences (typically expressed as ratings) on items that the users consume. From this information, recommender systems extrapolate user preferences for other available, not yet consumed, items, allowing the system to provide intelligent, personalized recommendations to individual users. In this way the system is able to recommend items with the highest predicted preferences/ratings, or to simply inform the user about the user's expected (i.e., system-predicted) preference for a given item, so that the users can make more informed consumption decisions.

However, as will be discussed in more detail below, recent research on recommender systems' biases (Adomavicius et al. 2013, 2018; Cosley et al. 2003) consistently demonstrates that interactions with a recommender system before consuming an item (and observing the system-predicted rating for that item) can significantly affect (i.e., contaminate) the users' self-reported post-consumption ratings and even the users' economic behavior, such as the users' willingness to pay. Of greater significance, observing online recommendations can bias users' subsequently self-reported preference ratings even *immediately after* the item consumption experience (Adomavicius et al. 2013). In research and practice, post-consumption user ratings are typically interpreted as an uncontaminated expression of the consumers' preferences. In other words, these ratings are not expected to be affected by the system recommendation either explicitly or implicitly, which is evidenced by the fact that they are routinely used by

the same recommender systems as the "ground truth" for further refinement of subsequent rating predictions. Thus, the observed post-consumption biases on ratings are problematic.

The unwanted post-consumption biasing of preferences can have a number of side-effects, such as affecting the system's future predictive performance due to deteriorating information quality (Adomavicius et al. 2013; Cosley et al. 2003). Biases can contaminate the recommender system's inputs, weakening the system's ability to provide high-quality recommendations in subsequent iterations; can pull consumers' preferences toward displayed recommendations, thus providing a distorted view of the system's performance; and can amplify the recommender system's vulnerability to shilling attacks by unscrupulous agents who inject false opinions to manipulate the system to operate in their favor. Therefore, an important and practical general research question is: What can be done to reduce (or prevent altogether) such recommendation biases/side-effects?

There are several potential ways to approach this question. As discussed later in the paper, one common strategy to address these types of issues is the "modify the decision environment" approach (Soll et al. 2016). In particular, we use a choice architecture strategy by changing the *information presentation* aspect of the decision environment (i.e., the manner in which information is presented/displayed to the decision maker). This approach has been advocated in the literature as a way to remove certain common biases (Thaler and Sunstein 2008). Specifically, in this paper, for a significant number of *rating interface displays* derived from those in common use and representing certain theoretical properties, we investigated their impact on generating (or moderating) recommendation biases. In other words, for this study, we operationalized the aforementioned generic research question as: What is the impact of features of common rating display designs on recommendation biases? A systematic test of different interface design components and their impact on recommendation biases has direct practical implications on the design and use of recommender systems.

In the rest of the introduction, we first delineate personalized recommender systems as the design task of interest. Next, we define and discuss bias within this context and highlight the potential problems that recommendation-induced biases can cause. The observed biases from prior research motivate our attempts to reduce bias in user ratings. After the introduction, two studies investigate the effects of different rating displays on the biases produced by personalized system recommendations.

Personalized Recommender Systems

A recommender system aims to suggest desirable products that consumers have not yet purchased, experienced, or considered (depending on the specific application context). Most recommender systems use consumer ratings of experienced items as inputs for the system's computational techniques that estimate personal preferences for items that have not yet been consumed by the individual. In this paper, we use the term *recommendation* to refer to any personalized, system-predicted rating value that is observed by the user (whether favorable or unfavorable). Real-world recommender systems often present the predicted user preference for an item as a system rating to indicate an expectation of how much a specific consumer will like that specific item,

These personalized ratings are distinct from other forms of commonly used quality-signaling information, such as aggregate user ratings (i.e., mean item ratings across all users). Although the aggregate user ratings may be structurally similar to personalized ratings (e.g., expressed on a similar 1 to 5 star scale), the former represent *non-personalized* item quality information and the underlying motivational theory for their use is very different. In particular, the value of aggregate ratings is typically argued in terms of social influence, a form of learning based on observing others (e.g., Salganik and Watts 2008). In contrast, personalized ratings typically are not computed based on any kind of social influence. For example, content-based algorithms match feature characteristics of items previously consumed by the same user, independent of the preferences of other users (Ricci et al. 2015). Even algorithms that incorporate preferences of other users, for example, collaborative filtering techniques (Konstan et al. 1997; Sarwar et al. 2001), generally do not involve an explicit, direct social component.

Also distinct from our approach are studies that examine outcomes at a market level (e.g., Fleder and Hosanagar 2009; Godinho de Matos et al. 2016). Our interest is in the effects of personalized rating interface displays on the expressed preferences of individual users. Whether the results generalize to the use of non-personalized aggregate quality information, or to market-level outcomes, or to other broader settings is not necessarily obvious based on possible differences in psychological processing or in the granularity of analysis, and are issues for further research.

Biases in Recommender Systems

Within the context of recommender systems, a few studies have shown strong and consistent evidence that ratings provided by consumers are biased toward system-generated recommendations. As the term *bias* is used in the literature and common discourse in different ways, it is important to be clear of the usage here. In most lay uses, bias is considered pejorative and representative of a negative prejudice. In behavioral economics and decision research, the word bias is used in an agnostic manner to represent a systematic pattern of deviation from a norm or rational standard of judgment (e.g., Haselton et al. 2015). Studies of cognitive biases examine how people make judgments and under which conditions those judgments are unreliable or violate normative expectations (e.g., Kahneman et al. 1982; Schkade and Johnson 1989; Thorsteinson et al. 2008).

As the operationalization of bias, we follow the original, related behavioral economic studies which measure the bias induced by the display context as the mean difference between people's judgments when a high versus low value is presented, stemming from the general principle of measuring bias as the difference in responses between conditions that should not differ according to the normative principles. For example, Tversky and Kahneman (1974) asked for the percentage of African nations in the United Nations with subjects given either a random high initial value (65) or a low initial value (10). The observation that those receiving a high value provided a higher judgment than those receiving a low value was indicative of an anchoring and (insufficient) adjustment bias. In our studies, we capture bias in a similar fashion: as a difference in judgments when users randomly receive higher versus lower system-predicted ratings. This approach has also been used in prior information systems literature on biases in recommender systems (Adomavicius et al. 2013).

As noted in this definition, a critical component of identifying bias is an arguably normative expectation. For personalized ratings, this argument hinges on an important issue of the *timing* of the user's judgment. A recommender system can provide predicted preference ratings that are useful to consumers prior to item choice and consumption. This is where the system has recognized value, and is the driver of the great popularity of personalized recommendations. But, after consuming or experiencing the good, this value is spent. Post consumption, the user is typically asked to provide a rating describing their actual preference for the experienced good. Since the consumer now has direct personal experience, his/her reported rating is presumed to provide an expression of the user's unadulterated preference for the consumed item. Furthermore, since the personalized recommendation no longer provides any useful information at this point, it is commonly (implicitly or explicitly) assumed that the user's reported rating is uncontaminated by the previously observed system's prediction, e.g., as highlighted by the fact that users' reported ratings are typically used to evaluate the recommender system's accuracy (by comparing how closely the system-predicted ratings match the users' reported ratings).
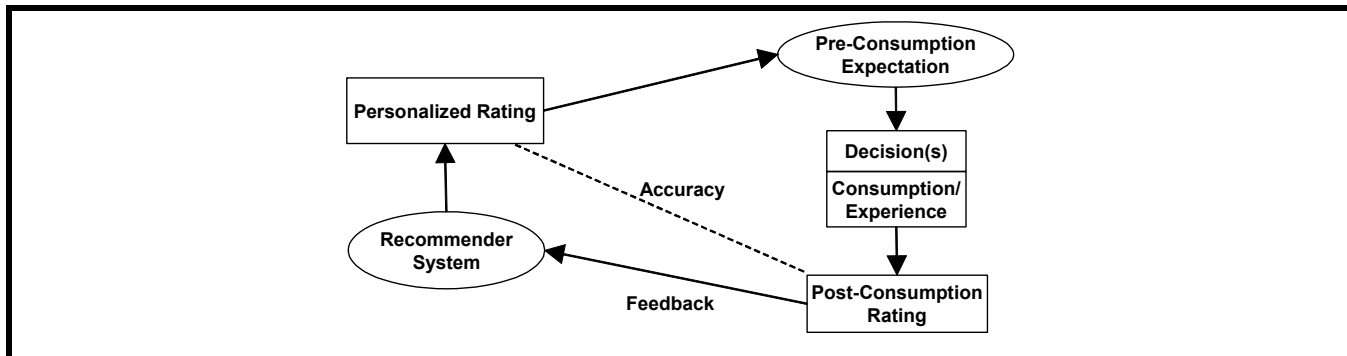
Figure 1. System Recommendations and User Ratings as Part of a Control System

Figure 1 shows the distinction as well as the potential consequences. The diagram shows the observable aspects of the situation in rectangles. At the pre-consumption stage (upper part of the figure) is the value proposition for the use of a recommender system, which presents personalized recommendations to users. The recommendation, along with any other information the user has, is used to form a pre-consumption expectation of the level of relevance and/or quality of the item under consideration. The consumer uses this information to make selection, purchase, and consumption decisions. This activity ends with the consumption and/or experience of the good or service. The effects of system recommendations on these selection and consumption decisions are not the focus of this research. After consumers experience items, they can submit their judgments of these newly consumed items in the form of item ratings back to the recommender system (the feedback arrow in Figure 1). As mentioned earlier, these post-consumption ratings are presumed to represent unadulterated preferences driven only by the user's experience of the item and not by the system's recommendation. A bias in users' judgments is present if the user ratings are still affected by the system's predictions, post item consumption. It is this bias that is of current interest. The user ratings are then used as input data by the system in future recommendation calculations on an ongoing basis. This completes a feedback loop as part of a control system that is central to a recommender system's future use and value.

As part of a feedback loop, the fidelity of users' post-consumption preference ratings significantly affects the system's ability to provide accurate and useful recommendations for consumers' future use. Consequently, recommendation-induced biases in post-consumption preference ratings can be harmful in several ways (Adomavicius et al. 2013; Cosley et al. 2003). Recommendation biases can distort or manipulate consumer preferences and purchasing behaviors, potentially leading to suboptimal product choices. Biases may allow third-party agents to manipulate the recom-

mender system so that it operates in their favor, which could reduce consumers' trust in the recommender system and harm its long-term value. Distorted user preference ratings that are submitted as inputs to recommender systems can potentially distort the judged accuracy of the system and reduce the system's effectiveness.

Empirically, biases from personalized recommendations have been well-established in prior work. For example, when users re-rated a movie while being shown a "predicted" personalized rating that was altered upward or downward from the original rating, users tended to give higher or lower ratings, respectively, as compared to a control group receiving accurate original ratings (Cosley et al. 2003). Thus, system predictions affected users' ratings based on recall, for movies seen in the past and now being evaluated. Adomavicius et al. (2013) further examined consumer preference ratings elicited at the time of item consumption, thus removing possible explanations deriving from the preference uncertainty that can be present at the point of recall (e.g., "How much did I like that movie last month?"). In their setting, no preference uncertainty was present that might encourage users to use the system recommendation to fill in the gaps in memory. Still, the displayed system-predicted ratings, when perturbed to be higher or lower, affected the submitted consumer ratings to move in the same direction. These biases have been consistently obtained across a variety of procedures and digital goods, including songs, TV shows, movies, and jokes (Adomavicius et al. 2013, 2018; Cosley et al. 2003;). In addition, the biases can occur both when the user observes the system's predicted rating for an item on some prior web page (before item consumption and before user's post-consumption rating collection) or when it is observed on the same page where the item is consumed and the rating is collected (Adomavicius et al. 2013). In other words, simply removing the system's recommendations from the rating collection page would not eliminate the bias. Also, these effects are substantial in their magnitude. For instance, if the recommendation that is observed before item consumption is perturbed by 1

star (on the 1 to 5 star scale), the user's self-reported post-consumption preference rating is shifted, on average, by 0.35 stars in the direction of biased recommendation. Going beyond the recommendation effects on preference ratings, Adomavicius et al. (2018) identify and measure a robust and strong effect of system recommendations on consumers' economic behavior, as measured by their willingness to pay. They observed marginal effects ranging from 7% to 17% in willingness to pay for 1-star changes in recommendations (again, on the 1 to 5 star scale), which reflects a practically significant level of bias. In other words, the bias introduced by recommendation systems is a robust, reliable, and significant effect on user's preference judgments.

In summary, although recommendations are of value in the pre-consumption phase, the normative expectation is that post-consumption preference ratings are uncontaminated by system recommendations. Evidence strongly indicates they are not. Therefore, understanding how these substantial biases in user judgments can be reduced in recommender systems represents an important and practical research question. We evaluate a user-interface-driven approach to reduce or remove bias in which we manipulate the format of the information presented in a personalized recommendation. We empirically test this approach using controlled experimentation with human participants.

Finally, substantial research literature already exists on various human decision biases as well as on the possibilities to reduce them, including some studies in the IS literature on certain biases in online product reviews (Hu et al. 2017; Li and Hitt 2008). At the same time, the existing literature also has many examples where findings do not automatically translate across different domains,[2] that is, the context plays a crucial role. Since modeling and prediction of *subjective user preferences* is one of the direct goals of personalized recommender systems, such systems represent a unique context that is largely underexplored from the de-biasing perspective.

Reducing Bias from Personalized Recommendations

Removing bias often is not a trivial endeavor. Soll et al. (2016) identify two general types of de-biasing strategies: *modify the person* and *modify the environment*. Strategies for modifying the person include pre-judgment interventions such as educating the decision maker about biases (e.g., Larrick

2004) and training decision makers on cognitive strategies that lead to less biased outcomes (e.g., Keeney 2012; Russo and Schoemaker 2014), or post-judgment interventions such as the use of decision models that take input from the decision maker and generate judgments for them (e.g., the APGAR test for evaluating the health of newborn babies; see Casey et al. 2001). In the context of our study, using this general approach has several difficulties. Pre-judgment modify-the-person strategies generally introduce costs to the user that would undermine the system's ability to obtain the needed inputs from the users. Also, such strategies rely on the decision maker acknowledging the judgment bias. Doing so in the context of personalized recommendations may undermine the user's trust in the system, impeding the system's ability to be useful (see, for example, Adomavicius et al. 2013). Post-judgment modify-the-person strategies have not been successful when applying algorithms based on aggregate biases (Adomavicius et al. 2014), and applying individualized algorithmic adjustments requires individualized susceptibility-to-bias information that is generally unavailable and difficult to attain.

Therefore, we focus on modify-the-environment approaches to overcome decision biases in recommender systems. Specifically, we adopt a choice architecture approach: changing the manner in which information is presented to the decision maker. Thaler and Sunstein (2008) advocate this approach as a way of removing common biases; a wide variety of choice architecture manipulations have been proposed and studied (for a review, see Soll et al. 2016). The literature has shown extensively that information format can significantly influence consumers' decision-making processes by affecting the cognitive costs and benefits of the tasks (e.g., Bettman and Kakkar 1977; Jarvenpaa 1989).

The interface displays that we test are selected based on a combination of practical and theoretical bases. Given the prevalence of recommender systems in the real world, whether personalized or otherwise (e.g., popularity-based), we made sure to include in the study the most popular, canonical rating scales used in practice. One canonical scale for presenting personalized recommendations is a 1 to 5 star rating scale shown graphically, either with or without the numeric prediction attached (e.g., see movie ratings on movielens.org or on Netflix's DVD.com). We test these two popular recommendation displays as the *Star-Only* and the *Star-Numeric* display treatments in our study.

Theoretically, evidence in communicating information about climate change has demonstrated the potential benefits of using a dual scale like the Star-Numeric display for promoting consistent communication (Budescu et al. 2012). Alternatively, adding the numerical value to the graphical interface may add scale compatibility effects that increase the level of

---

[2]For example, while anchoring-type effects have been observed in some application domains, they are not observed (or are more modest) in other domains. For examples, see Maniadis et al. (2014), Simonson and Drolet (2004), and Tufano (2010).

bias, thus favoring the Star-Only display. Scale compatibility effects occur when there is a correspondence between the stimulus and response scales (e.g., system-predicted recommendations and user-provided preference ratings are both reflected on the same numeric 1–5 scale), leading to the stimulus biasing the response in the decision process (Tversky et al. 1988). Breaking up this correspondence is a de-biasing strategy that has shown promise in other settings (e.g., Burson et al. 2009; Camilleri and Larrick 2014; Larrick and Soll 2008). Study 2 further investigates the role of scale compatibility as it relates to the collection of consumer ratings and display of recommendations.

A third interface for providing ratings that is common in practice is a binary design used to denote high versus low predictions. This binary recommendation (a thumb up or down associated with the item) is the simplest format possible for communicating personalized product recommendations of items, and can be seen in use by many online music services such as Apple's iTunes, Pandora, and Spotify. One aspect that is introduced by this display is that of vagueness. Prior research has provided several exemplar domains where less certain estimates presented as imprecise values, such as range forecasts compared to point forecasts, have been advantageous for removing biases, including for policy decisions (e.g., Dieckmann et al. 2010), in the use of weather forecasts (Joslyn and LeClerc 2012), and for professionals making earnings forecasts (Du et al. 2011).

Based on the key identified features associated with these three common display formats and their high-level features (e.g., graphical versus numeric and precise versus vague representation), we identify four additional generic/prototypical rating display formats for testing. These displays extract the two defining high-level features and manipulate them within a 2 × 2 factorial combination of two levels each for the two dimensions of interest: graphical versus numerical representations and precise versus vague representations. Adding these four display formats, Table 1 summarizes the seven rating representation options (i.e., *Numeric-Precise*, *Numeric-Vague*, *Graphic-Precise*, and *Graphic-Vague* as generic display formats and *Star-Numeric*, *Star-Only*, and *Binary* as real-world display formats) with examples of each.

Including the four generic rating display formats in our study allows for a more systematic investigation of potential theoretical underpinnings along the two dimensions of interest. In particular, prior research has shown that graphic and tabular (that are primarily numerical) representations emphasize different characteristics of a given data set and, therefore, are better suited for different tasks (Vessey 1991) and, thus, may be anticipated to have differing impacts on user bias. Furthermore, studying the impact of graphical versus numerical presentations of data is also supported by a long Information

Systems research history on graphs and tables (e.g., Braithwaite and Goldstone 2013; Meyer et al. 1999; Porat et al. 2009), the focus of which has been on understanding whether graphical or text-based presentations of data performed better in aiding decision makers, particularly in identifying relationships among variables. The results in these studies are inconsistent, and there is good evidence that the specific context and decision-making task play important roles. Also, to the best of our knowledge, the prior literature on graph versus textual representations has not considered subjective preference judgements; thus, the current study represents a contribution in this regard.

We also note that, as shown for the graphical displays in Table 1, we move away from the more typical star rating displays commonly used in practice. We did this to remove a potentially confounding aspect of the star displays that occurs when displaying fractional predictions (e.g., a rating of 3.2). Typically, a fractional rating is displayed as a shading of an imagined box within which the star fits, which results in improper scaling. The bar-shape graphical displays alleviate this issue and, although less popular than star displays, can be found in use by many real-world websites, including Orbitz, Kayak, and Rotten Tomatoes.[3]

The commonly used binary scale can be considered a maximally vague recommendation, indicating only whether the predicted rating is above or below some cutoff. With respect to the precision versus vagueness factor, we generalize from the binary scale and examine the possible effects of more moderate levels of vagueness as compared to precise predictions—for example, whether the predicted preference rating is in the *range* 3.4–4.2 as opposed to a specific predicted preference rating value of 3.8–upon the user's preference judgments post consumption. As mentioned earlier, range-based (as compared to precise, point-based) information presentation has been advantageous for removing biases in some application domains (e.g., for weather forecasts). One plausible explanation for this effect is that imprecision prompts the consideration of alternative possibilities, a strategy that has proved effective in reducing a variety of judgmental biases (see Hirt et al. 2004). Another, complementary explanation is suggested by Hsee (1995), in which including range information about an attribute (e.g., "the bird population will range from 100–500") helped prevent decision makers from overweighting the attribute in their decision. Overall, the prior studies on imprecision have largely focused on tasks having verifiable outcomes in the future (e.g., whether an investment will actually return X% in the specified future time period). The present context is quite different in that the system is predicting a subjective preference that is

---

[3]Some examples are provided in Appendix D.

Table 1. Example Displays of System Predicted Ratings (Between-Subjects Conditions)		
Group	**N**	**Example Display of Predicted Rating**
Numeric-Precise	40	3.0 (out of 5)
Numeric-Vague	39	between 2.6 and 3.4 (out of 5)
Graphic-Precise	40	Hate it [graphic bar] Love it
Graphic-Vague	40	Hate it [graphic bar] Love it
Star-Numeric	45	★★★☆☆ 3.0 (out of 5)
Star-Only	43	★★★☆☆
Binary	40	👍 Thumb Up or 👎 Thumb Down

not externally verifiable. This extension of prior work to the realm of preferences is novel to the current area of research. Specifically, in many commercial recommender systems, the predicted preference estimates (i.e., system recommendations) are presented as precise and exact values, typically on a scale from 1 to 5 with one decimal point precision (e.g., 4.3 out of 5 stars). If similar uncertainty dynamics apply in the context of providing system recommendations to aid preference decisions, then rating display formats with decreased precision can potentially lower the bias induced by typical system recommendations.

Together, the seven rating display formats that we examine in this research provide representative coverage of key possible rating display components. We compare them to address the central question of interest: *Does the level of bias introduced by personalized recommendations vary depending on the characteristics of the recommendation display format?*

Study 1: De-Biasing Through Various Rating Display Types

Methodology

The first study involved recommendations and ratings of jokes, so the participant population required no special characteristics. Jokes are a stimulus that can be experienced in the lab session so that the readers' preference ratings can be gathered immediately after reading the joke; there is no uncertainty of preference due to recall effects. The standard assumption in such a situation is that the participant's reported rating should provide an unadulterated expression of their preference. Jokes are also stimuli that have elicited

biases at a similar rate as movies, TV shows, and songs (see Adomavicius et al. 2013, 2016, 2018; Cosley et al. 2003), making them a good setting for investigating de-biasing efforts.

Subjects

The study was conducted at a behavioral research lab, and participants were recruited from a college's research participant pool. In total, 287 people completed the study. 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 Table 2.

Procedure

As noted in the previous section and exemplified in Table 1, seven display formats were used in the study: Numeric-Precise, Numeric-Vague, Graphic-Precise, Graphic-Vague (forming a 2 × 2 design), Star-Only, Star-Numeric, and Binary. For the two vague prediction conditions, we used an interval of +/- 0.4 for the possible values for the personalized recommendation rating (i.e., if the personalized rating was 4.1, the *Vague* conditions would represent it as range [3.7, 4.5] to the user). The seven conditions comprise a between-subjects set of randomized manipulations; the number of respondents in each condition is provided in Table 1.

A database of 100 jokes was used during the study. The jokes and the user rating data for training the recommendation algorithm were taken from the Jester Online Joke Recom-

Table 2. Participant Summary Statistics of Study 1	
Number of participants	287
Age: Mean (SD)	22.7 (4.68)
Gender	144 M, 143 F
Native speaker of English	70.7%
Prior experience with recommender systems	74.9%
Student level	185 undergraduate, 87 graduate, 15 others

mender System repository (Goldberg et al. 2001; http://eigentaste.berkeley.edu/dataset). Beginning with Jester Dataset 2 (150 jokes), we successively removed (1) jokes that were suggested for removal at the Jester website (e.g., those with insufficient user ratings); (2) jokes that more than one of our coauthors identified as having potentially objectionable content; and (3) jokes that were greatest in length (based on word count). The order of the jokes was randomized across participants.

Upon logging on to their experiment workstation, participants were randomly assigned to one of the seven rating display treatment groups (see Table 1). All recommendations presented to participants were in the format of their rating display group. The experimental procedure consisted of three tasks, all of which were performed using a web-based application on personal computers with dividers providing privacy between participants.

Task 1. Each participant was asked to read 50 jokes randomly selected from the pool of 100 jokes and provide their preference (i.e., post-consumption) ratings for them. For each joke, we also asked participants to indicate whether they had heard the joke before (Yes/No radio buttons). For the preference ratings, we instructed participants to rate the jokes using a five-star scale with half-star responses allowed. The scale values were described as * = Hate it, ** = Don't like it, *** = Like it, **** = Really like it, and ***** = Love it.[4] Participants responded by selecting a rating value from a drop-down box, using only the numerical values (1, 1.5, 2, …, 5) and the verbal labels (on the whole number values, as defined above) in the response area. Users submitted all preference judgments in this study using this scale, regardless of which of the seven personalized recommendation format treatment conditions a participant is assigned, so that the effect of varying the system display could be isolated.

---

[4]We note that this scale is asymmetric; the middle rating on the 1–5 scale is 3, which has a positive interpretation. However, this is the most commonly used scale in practice (e.g., Amazon, Barnes and Noble, Yelp and MovieLens), and the focus of our study and analysis is on relative effects, not absolute effects.

Task 2. From the remaining unrated and unseen 50 jokes, participants in all seven experimental groups read 30 randomly selected jokes in six within-subjects conditions (five recommendation conditions and control) as summarized in Table 3—five jokes per condition.

The first two recommendation conditions used randomly generated predicted ratings that were not based on user preferences or any real system-generated recommendations (i.e., artificial recommendations). These conditions allow us to study the effects of system predictions in a somewhat pure form, albeit removed from the reality of tying the system predictions to the user's preferences in any way. Using random recommended values, any effect is due to the differences in the displays themselves. Since multiple items were presented to the participant in each condition, the artificial recommendations were randomly drawn from uniform distributions so that the same recommendation values were not repeatedly presented for all items. In the *High-Artificial* condition, we presented randomly generated high recommendations between 3.5 and 4.5 stars (drawn from a uniform distribution). In the *Low-Artificial* condition, we used randomly generated low recommendations between 1.5 and 2.5 stars (drawn from a uniform distribution).

For the next three conditions, real system-generated recommendations were used. Using the system-generated recommendations extends the approach of using artificial recommendations and provides a convergent validity to our results. The ratings provided in Task 1 were used as seed information to calculate *personalized* predicted preference ratings (i.e., recommendations) for the jokes in these three conditions. Recommendations were calculated using state of the art recommender system algorithms, commonly used in practice and which we empirically validated (see Appendix C for details). By perturbing actual system predictions, we control for the differing preferences of the users up front, providing a more realistic setup in which the perturbations reflect errors in the system's predictions (in this case, systematic errors that we introduce). In the *High-Perturbed* condition, the real algorithmic predictions were perturbed upward by 1 star. In the *Low-Perturbed* condition, they were perturbed downward

Table 3. Experimental Conditions for the Displayed System-Predicted Ratings (manipulations are within-subjects)

Group	N	Description
High-Artificial	5	Randomly generated high recommendations, ~U[3.5, 4.5]
Low-Artificial	5	Randomly generated low recommendations, ~U[1.5, 2.5]
High-Perturbed	5	Actual predictions that were perturbed upward by 1 star
Low-Perturbed	5	Actual predictions that were perturbed downward by 1 star
Accurate	5	Actual algorithmic predictions (i.e., not perturbed)
Control	5	No predictions were provided

by 1 star. In the *Accurate* condition, the actual algorithmic predictions (i.e., not perturbed) were presented. Finally, the *Control* condition included no recommendation ratings.

For all jokes and participants, participants responded by providing their post-consumption ratings using the same preference scale and drop-down box that was used in Task 1 and described above. In all five recommendation conditions, the rating response options were displayed on the same page as a given item (joke) and the system's recommendation. This represents a realistic, expected scenario to the users—many real-world recommender/rating systems (e.g., Netflix, Rotten Tomatoes, IMDB; screenshots are provided in Appendix D) present the system ratings on the same page as the user's rating collection interface. Furthermore, as mentioned earlier, prior research has robustly demonstrated the existence of recommendation biases when the user observes the system's recommendation for an item either on the same page where the rating is collected or on some prior web page (before item consumption and before rating collection). The goal of this paper is to ascertain comparative advantages/disadvantages of different rating display characteristics with respect to recommendation biases, while holding constant across all treatment conditions the "when" and "where" system rating information is shown with respect to each joke.

The five recommendation conditions along with the Control condition (see Table 3) formed a *within-subjects* experimental manipulation; the seven different interface displays (Table 1) formed a *between-subjects* manipulation applied for each of the five recommendation conditions (excluding the control where no recommendation was given). The artificial ratings and the perturbations of the system ratings allow us to test for bias in user preferences due to recommendations, measured as a difference between high and low manipulations (within-subjects). By randomly assigning participants to different interface groups, we can test to see if certain interface designs eliminate or change the magnitude of the recommendation bias, measured as changes in the high/low difference across interface groups. Therefore, the within-subjects manipulation is for establishing the existence of biases due to recommen-

dations, and the between-subjects manipulation is for comparisons of biases across interface designs.

For the within-subjects component of the experiment, we first selected five jokes for the High-Perturbed condition and five jokes for the Low-Perturbed condition. These 10 jokes were chosen pseudo-randomly to assure that the manipulated ratings would fit into the 5-point rating scale (i.e., the predicted rating plus/minus 1 star perturbation must be a number between 1 and 5). Among the remaining jokes, we randomly selected 20 jokes and assigned them to the four other groups: 5 to Accurate, 5 to High-Artificial, 5 to Low-Artificial, and 5 as a control with no recommendation rating provided. The 25 jokes with recommendations were mixed, randomly ordered, and presented on five consecutive webpages (with 5 jokes displayed on each page). The five control jokes were presented on the subsequent webpage with instructions indicating that these were jokes for which "our system did not have enough data to make predictions," thus providing an explanation for why no recommendations accompanied these items.

Task 3. Participants completed a short survey that collected demographic and other individual information for use in the analyses (see Tables 2 and 5).

The Perturbed versus Artificial within-subjects manipulations represent two different approaches to the study of recommendation system bias. In this paper, we report the results of the analysis for the Artificial recommendations only. The Artificial recommendation manipulations provide a view of bias that controls for the value ranges shown, manipulating some to be high and some low, but they do not account for individual differences in preferences. However, we use a *post hoc* measure to provide a control for such individual differences. The Perturbed recommendation manipulations explicitly control for possible preference differences within the experimental design. The results are comparable and consistent across the two approaches, and the results for the Perturbed recommendation manipulations are provided in Appendix A.

Experimental Results: Artificial Recommendations in Various Display Types

As a check of the participant randomization done in our study, we compared the mean submitted post-consumption ratings for the jokes rated prior to treatment and for the control jokes viewed after treatment across the seven treatment groups. All of these jokes are ones for which the participants did not see a system recommendation. ANOVA analyses support the conclusion that proper randomization was performed ($p > .30$ for both tests), and any effects observed are driven by the experimental manipulations. As another check, we calculated the average time subjects spent on each page across all seven experiment groups. We then performed *t*-tests to compare the average time between different groups, and did not find any significant differences. In other words, all seven rating interfaces required approximately the same amount of time for subjects to complete the rating task (i.e., they did not differ significantly in the amount of effort required).

We begin the investigation of the manipulations with the analyses of all the seven rating display conditions in our experiment. Figure 2 presents a plot of the aggregate means of user-submitted post-consumption ratings for each of the treatment groups when high and low artificial recommendations were provided, as well as for the control stimuli for each group where no recommendations were provided. For each rating display format, Table 4 shows the pair-wise *t*-tests to compare user-submitted ratings after receiving high and low artificial recommendations and no recommendation (control).

If no bias existed, we would expect there to be no significant difference among the high, low, and control conditions. Instead, from the figure and table, we observe that low (high) artificial recommendations pull down (up) user's post-consumption preference ratings relative to the control. All comparisons between high and low conditions are significant across the seven rating representations (one-tailed $p < 0.001$ for all high versus low tests), showing a clear, positive effect of randomly generated recommendations on consumers' preference ratings. All effect sizes are large (Cohen's *d* values range between 0.71 and 1.23). The control condition demonstrated intermediate preference ratings, showing a significant difference from both the High and Low conditions for the majority of the rating display options. In other words, we found that none of the seven rating display options could completely remove the biases generated by recommendations.

However, not all interfaces were equivalent in terms of their bias reduction performance. Among the seven rating representations, Numeric-Precise demonstrated the largest difference in consumer ratings: a 0.793 difference between high and low conditions on a scale from 1 to 5. In other words, the

observed level of recommendation bias is substantial: about 20% of the entire rating scale. Binary showed the smallest resulting difference: 0.408. This shows that bias reduction can be substantial as well; up to half of the bias can be eliminated by choosing a different rating display interface. Such results suggest that various rating displays have differential effects on consumers' preference ratings. To check this globally, we computed rating differences between high and low conditions for each participant and performed a one-way ANOVA across displays. We see a significant difference in effect sizes among different rating representations ($F(6, 280) = 2.24$, $p < 0.05$). Since the overall effect was significant, we performed regression analyses to explore the differences in bias between different rating displays, while controlling for participant-level factors. We excluded from our analyses a small fraction (0.86%) of cases in which participants reported having previously seen the joke. The reason for focusing only on the first-time consumption cases is that the participants' preferences for previously consumed jokes are more likely to involve other confounding factors (recall, prior consumption context, etc.) that might affect the subjects' judgements of these jokes.

The repeated-measures design of the experiment, wherein each participant was exposed to both high and low artificial recommendations in a random fashion, allows us to model the aggregate relationship between shown ratings and users' submitted post-consumption ratings while controlling for individual participant differences. We apply a random effects generalized least squares (GLS) model using robust standard errors, clustered by participant, and using participant-level controls for the analysis:

$$
\begin{aligned}
UserRating_{ij} = \ & b_0 + b_1(High_{ij}) + b_2(Display_i) \\
& + b_3(Display_i \times High_{ij}) \\
& + b_4(ShownRatingNoise_{ij}) \\
& + b_5(PredictedRating_{ij}) \\
& + b_6(AdditionalControls) \ + u_i + \varepsilon_{ij}
\end{aligned}
$$

In the model, $UserRating_{ij}$ is the submitted rating for participant i on joke j. $High_{ij}$ indicates whether the shown rating for participant i on joke j is a high or low artificial recommendation. Thus, the coefficient on $High_{ij}$ measures the difference in user post-consumption preference ratings between high and low conditions (manipulated within-subjects), which is our operationalization of bias. Since we are not interested in detecting asymmetries in bias and this measure is the most sensitive to any effects that may be occurring, it is an appropriate measure of bias for this analysis. $Display_i$ is the rating display option shown to participant i. Because $Display_i$ is a nominal variable with seven levels, this variable entered into the model as six indicator variables. Consequently, the interaction term $Display_i \times High_{ij}$ captures the dif-
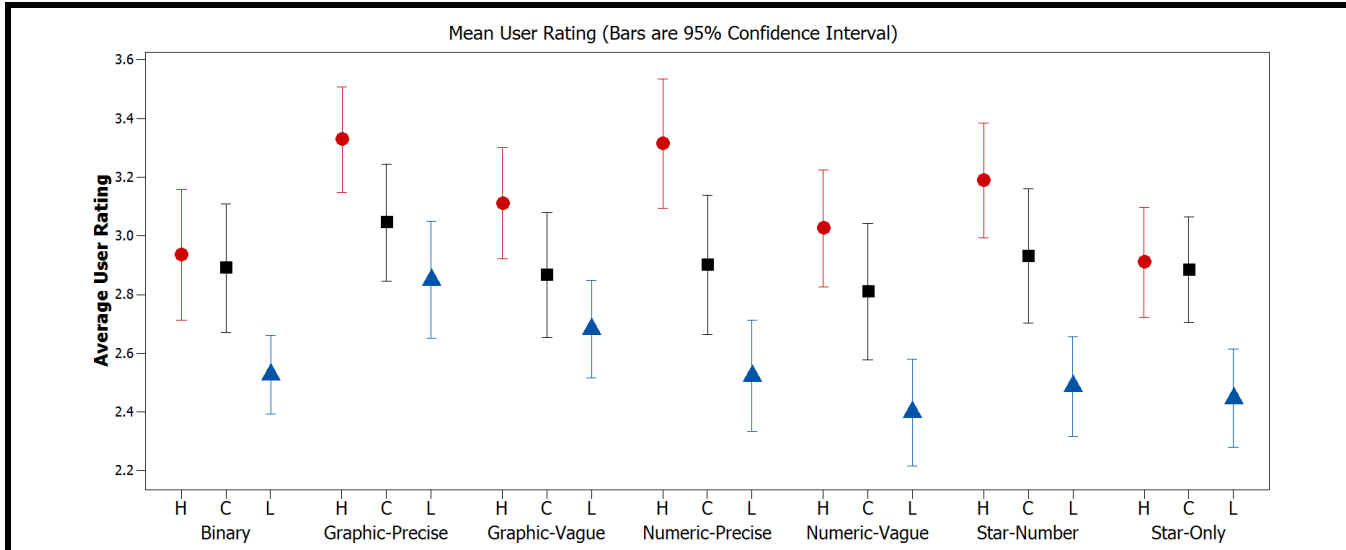
Figure 2. Means and 95% Confidence Intervals of User Submitted Ratings After Receiving High Artificial (High: dot), Low Artificial (Low: triangle), or No Recommendations (Control: square)

Table 4. Pair-Wise Comparisons of Mean User Rating Difference for Each Rating Display Using *T*-Tests			
Rating Display	**High – Low**	**High – Control**	**Low – Control**
Binary	0.408***	0.045	-0.363***
Graphic-Precise	0.478***	0.283**	-0.195*
Graphic-Vague	0.428***	0.245**	-0.183*
Numeric-Precise	0.793***	0.415***	-0.378***
Numeric-Vague	0.628***	0.215*	-0.413***
Star-Numeric	0.702***	0.258***	-0.444***
Star-Only	0.463***	0.026	-0.437***

*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$

ference in bias across different displays. Our initial regression analysis uses the common Numeric-Precise rating display condition as the baseline rating representation to compare with the other six options.

Recall that we introduced randomness into the values for the artificial high and low recommendation conditions by drawing from uniform distributions: High [3.5, 4.5] and Low [1.5, 2.5]. So, for example, for a joke in the High Artificial condition, the participant saw a predicted rating of $4.0 + \delta$, where δ was a random noise factor ranging between -0.5 and 0.5 (both inclusive). We define $ShownRatingNoise_{ij}$ as a derived variable that captures this noise factor δ that varied for each participant and for each joke in the artificial conditions. The value for $ShownRatingNoise_{ij}$ is computed by either subtracting 4.0 from the shown rating (High artificial condition) or 2.0 from the shown rating (Low artificial condition).

$PredictedRating_{ij}$ is the predicted recommendation for participant i on joke j, calculated after the fact. Recall that the artificial rating seen by the participant did not depend on this value at all; however, the inclusion of this term provides a control for differing expected preferences among participants. $AdditionalControls$ is a vector of joke- and participant-related variables. The controls included in the model were the joke's funniness (average joke rating in the Jester dataset, continuous between 1 and 5), participant age (integer), gender (binary), school level (undergrad yes/no binary), whether they are native speakers of English (yes/no binary), whether they have prior experience with recommendation systems (yes/no binary), and whether they thought the recommendations in the study were accurate (interval five-point scale) and useful (interval five-point scale). The study utilized a repeated measures design with a balanced number of observations on each participant. To control for participant-level heterogeneity, the

composite error term ($u_i + \varepsilon_{ij}$) includes the individual participant effect u_i and the standard disturbance term ε_{ij}. A random effect model is used for participant heterogeneity since these individual-specific effects are uncorrelated with the randomly applied treatment conditions.

Table 5 indicates that randomly generated recommendations displayed in Numeric-Precise format can substantially affect consumers' preference ratings, leading to significant coefficients for *High* and *ShownRatingNoise*. The magnitude of the biases is substantial and consistent with that observed in prior research. For the Numeric-Precise group that has been studied previously, we observe a High coefficient of 0.793 for high versus low conditions. As the mean system rating difference between the high and low conditions is 2 on average, we see a bias effect of about 0.4 for every 1-star difference in the predicted rating.

More importantly, we found significant negative interaction effects between several rating display options and *High*. This indicates that Binary, Graphic-Precise, Graphic-Vague, and Star-Only displays, when compared to Numeric-Precise, can generate significantly lower biases in consumers' preference ratings. All the corresponding interaction terms have negative coefficients with *p*-values < 0.05. The coefficients on the interaction terms (ranging between -0.329 and -0.426) indicate a reduction in the bias on the order of 40% to 50% compared to the baseline Numeric-Precise case. However, the interaction terms for Numeric-Vague and Star-Numeric were not significant, hence no effect was suggested for these two display options as compared to the level of bias for the Numeric-Precise display.

To obtain a comprehensive view of how the seven rating representations differ from each other, we performed separate regression analyses with each of the seven rating displays used as the baseline in the same random effects GLS regression model. The results (omitted due to space constraints), exhibit consistent findings in that (1) none of the seven rating representations completely eliminate recommendation-induced biases in consumers' post-consumption preference ratings, (2) a Numeric-Precise format significantly increases biases relative to any of the non-numeric formats, and (3) a binary format significantly reduces biases compared to any of the numeric formats.

Thus, these initial analyses suggest a numeric-graphic display difference that we can pursue directly within a regression analysis that focuses on the four rating displays comprising the 2 × 2 factorial design. Recall that the first factor is the presentation style (Numeric/Graphic), and the second factor is the precision of the recommendation (Vague/Precise). We again applied a random effect GLS model using robust stan-

dard errors, clustered by participant, and with participant-level controls for the analysis:

$$
\begin{aligned}
UserRating_{ij} = \quad & b_0 + b_1(High_{ij}) + b_2(Presentation_i) \\
& + b_3(Precision_i) \\
& + b_4(Presentation_i \times Precision_i) \\
& + b_5(Presentation_i \times High_{ij}) \\
& + b_6(Precision_i \times High_{ij}) \\
& + b_7(ShownRatingNoise_{ij}) \\
& + b_8(PredictedRating_{ij}) \\
& + b_9(AdditionalControls) + u_i + \varepsilon_{ij}
\end{aligned}
$$

All variables are defined as in the previous model, with $UserRating_{ij}$ being the submitted rating for participant i on joke j. $Presentation_i$ and $Precision_i$ are added binary variables that indicate the rating display option shown to participant i. $Presentation_i$ is coded to be 1 if it is Numeric and 0 if it is Graphic. $Precision_i$ is coded to be 1 if it is Precise and 0 if it is Vague. $High_{ij}$ indicates whether the shown rating for participant i on joke j is a high or low artificial recommendation. The interaction terms $Presentation_i \times High_{ij}$ and $Precision_i \times High_{ij}$ capture the differences in bias between the levels of the two factors in the model.

The results of the regression analysis on these four conditions are presented in Table 6. In terms of effect size for the baseline, commonly studied Numeric-Precise condition, we see the usual effect magnitude. Combining the relevant coefficients, we observe a coefficient of 0.404 + 0.254 + 0.104 = 0.762 for High versus Low conditions in the Numeric-Precise group. Because the mean system rating difference between the High condition and the Low condition is 2 on average, we see a bias effect of almost 0.4 for every 1-star difference in the predicted rating.

Turning to the experimental treatment variables, we see that the Precision main effect is not significant ($b = 0.102$, $p = 0.150$). More importantly, the interaction between recommendation precision and recommendation direction (i.e., Precision × High, $b = 0.104$, $p = 0.148$) was not statistically significant. This term captures the change in the High/Low difference (i.e., bias) between the precision levels. The lack of support for an effect can also be observed visually in Figure 3b, where the slopes are not detectably different. Recommendations displayed in precise forms and vague forms have a similar level of effect in influencing user ratings.

In contrast, the presentation factor (numeric versus graphic) is significant ($b = -.264$, $p = 0.0035$). More importantly, the interaction between information representation and the direction of the artificial rating (high or low) was significant (i.e., Presentation × High, $b = 0.254$, $p = 0.005$). Figure 3a indicates the difference between ratings for High and Low

Table 5. Regression Analysis on High and Low Artificial Recommendations (Baseline: Numeric-Precise; Dependent Variable: UserRating)

	Coefficient Estimate	Standard Error
High	0.793***	(0.107)
ShownRatingNoise	0.287***	(0.059)
PredictedRating	0.277***	(0.036)
Display		
Numeric-Vague	-0.068	(0.114)
Star-Numeric	-0.025	(0.114)
Star-Only	-0.033	(0.111)
Graphic-Precise	0.301*	(0.119)
Graphic-Vague	0.169	(0.108)
Binary	0.036	(0.103)
Interactions		
Numeric-Vague×High	-0.169	(0.140)
Star-Numeric×High	-0.126	(0.146)
Star-Only×High	-0.344*	(0.140)
Graphic-Precise×High	-0.329*	(0.151)
Graphic-Vague×High	-0.363*	(0.143)
Binary×High	-0.426**	(0.142)
Controls		
jokeFunniness	0.594***	(0.070)
Age	-0.005	(0.005)
Male	0.057	(0.042)
Undergrad	-0.131*	(0.063)
native	-0.071	(0.053)
IfUsedRecSys	0.041	(0.052)
PredictionAccurate	0.068**	(0.025)
PredictionUseful	0.030	(0.018)
Constant	-0.474	(0.287)
R^2 within-subject	0.150	
R^2 between-subject	0.551	
R^2 overall	0.246	
X^2	770.27***	

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

| Table 6. Regression Analysis on Artificial Recommendations, for Numeric/Graphic and Precise/Vague Rating Displays (Dependent Variable: UserRating)	Coefficient Estimate	Standard Error
High	0.404***	(0.083)
ShownRatingNoise	0.203**	(0.077)
PredictedRating	0.224***	(0.042)
Presentation (Numeric = 1, Graphic = 0)	-0.264**	(0.098)
Precision (Precise = 1, Vague = 0)	0.102	(0.098)
Presentation×Precision	0.002	(0.118)
Presentation×High	0.254**	(0.099)
Precision×High	0.104	(0.099)
Controls		
jokeFunniness	0.712***	(0.084)
Age	-0.008	(0.007)
Male	0.062	(0.059)
Undergrad	-0.189*	(0.084)
native	-0.121*	(0.071)
IfUsedRecSys	0.079	(0.071)
PredictionAccurate	0.078*	(0.036)
PredictionUseful	0.027	(0.025)
Constant	-0.438	(0.382)
R^2 within-subject	0.166	
R^2 between-subject	0.516	
R^2 overall	0.252	
X^2	415.39***	

$*p \leq 0.05, **p \leq 0.01, ***p \leq 0.001$

conditions (slope of the line) is greater in the numeric conditions compared to the graphical conditions. Comparing coefficients, we observe a reduction of bias of about 1/3 by using graphical compared to numerical displays.

Thus, the results of the 2×2 factorial design support that personalized recommendations presented in a numeric format create larger bias in reported preferences than recommendations presented in a graphical format. Randomly assigned recommendations presented in any non-numeric format (including Binary, Graphic-Precise, Graphic-Vague, and Star-Only) generate smaller biases compared to the same recommendations displayed in the numeric formats (Numeric-Precise, Numeric-Vague, and Star-Numeric). No differences between vague and precise displays are supported. The display format of recommendations (e.g., numeric versus non-numeric) appears to be the dominant factor that determines the size of bias in consumers' post-consumption preferences.

Study 2: The Interplay Between Recommendation Display and Rating Response Interfaces

In Study 1, we manipulated the format in which the recommendation was presented to the user. However, in all cases the users submitted their post-consumption preference ratings as numerical responses on a 1–5 scale. By manipulating the format of the *user responses* in addition to the format of the system recommendations, we can further our understanding of the effects of numerical and graphical displays on response bias. The motivation for this study partly arises from the earlier theoretical discussion hypothesizing a graphical/numerical difference in bias based on a scale compatibility effect between the system's prediction and the consumer's post-consumption response. In other words, biases often arise when the presentation scale and response scale are highly
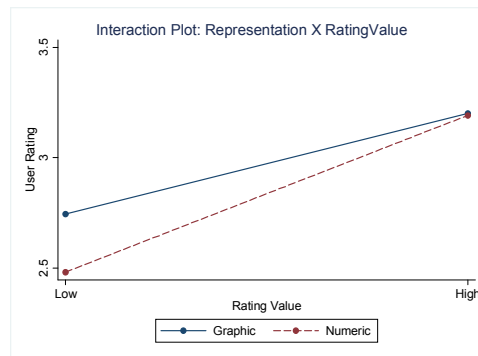
(a) **Information Representation (Numeric or Graphic) and Artificial Rating Value (High or Low)**
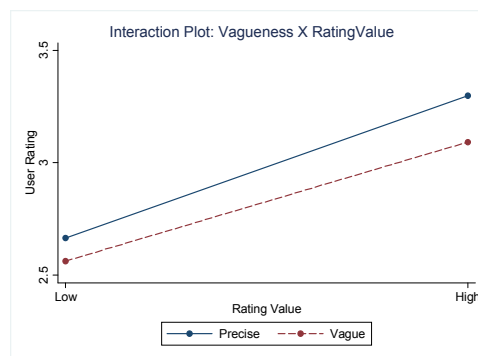
(b) **Precision of Recommendation (Precise or Vague) and Artificial Rating Value (High or Low)**

Note: The dotted reference line denotes the mean uer pre-treatment ratings.

Figure 3. Interaction Plots of Rating Displays Forms and Rating Values

similar (e.g., system-predicted recommendations and user-provided preference ratings are both measured on the same numeric 1–5 scale), leading to the stimulus being given greater weight in the decision process (Tversky et al. 1988). The ensuing consequence is that similarity in scales between the rating display and the response can magnify the bias (Tversky and Kahneman 1974). The scale compatibility hypothesis has received strong support in prior literature on judgment and choice (e.g., Fischer and Hawkins 1993; Slovic et al. 1990).

Prior studies also have suggested that the required effort is reduced when there is a close match between the problem representation and the problem-solving processes required to resolve the decision problem (Vessey 1991). The information (i.e., system-predicted recommendation) and task (i.e., provide a user rating) systematically interact to produce "compatibility" effects. Consequently, the information metrics that match the problem-solving processes required to

form a preference will have the greatest influence on responses (Fischer and Hawkins 1993).

In the recommender systems literature, Adomavicius et al. (2018) provided evidence that system predictions create bias even when scale compatibility was not present between the prediction and the preference response; however, their experiments were not designed to detect possible debiasing effects arising from breaking the scale compatibility between predicted ratings and responses. In Study 2, we explicitly test the difference in bias based on a scale compatibility effect between the system's prediction and the consumer's post-consumption response.

To the extent that scale compatibility effects are driving the numerical versus graphical difference observed in Study 1, the bias should be greater when the response format (either graphical or numeric) *matches* the rating display format. By varying the response format in addition to the predicted rating

format, we can test this explanation as well as explore an alternative interface approach for reducing bias. Thus, we conjecture that, when users provide preference ratings in a response format that is the same as the format of presented system recommendations (graphical versus numerical), greater bias in preference judgments will result as compared to when the two are in different formats.

Methodology

In total, 164 people completed Study 2. Demographic features of the sample are summarized in Table 7. The procedure of Study 2 was identical to that of Study 1, with each subject completing the same three tasks. The difference between Studies 1 and 2 lies in the between-subject experimental groups employed in Task 2. Study 2 uses a 2 × 2 between-subjects design to manipulate the post-consumption rating response interface and recommendation display interface with the following treatment conditions: *GraphicDisplay-GraphicResponse, GraphicDisplay-NumericResponse, NumericDisplay-GraphicResponse,* and *NumericDisplay-NumericResponse*. Table 8 presents examples of the rating display and response interfaces for the four groups. In the NumericResponse groups, participants were asked to rate the items using a drop-down box which included a 5-point rating scale with half-point increments (i.e., using the same response format as in Study 1). In GraphicResponse groups, participants were given a graphic rating canvas with color intensity indicating the level of preference and were instructed to click anywhere in the canvas area to give ratings. Multiple clicks were allowed for participants to adjust their ratings, and the last click for each item was recorded as the submitted rating. As in Study 1, in all four experimental groups the presented recommendations were displayed adjacent to the rating response options.

Experimental Results: Interface Compatibility

Following the analyses of Study 1, we compared the mean user post-consumption ratings of the four experimental groups after receiving *artificial* recommendations. As in Study 1, the results for the *perturbed* recommendations are consistent (except as indicated below), provide a complementary view, and are described in Appendix B for comparison. Table 9 summarizes the aggregate means. All High versus Low comparisons are significant across the four experimental conditions, showing a significant bias for all display-response combinations. As with Study 1, none of the interface designs were able to remove the bias completely.

We follow the same regression procedure as in Study 1, creating a panel from the data. The random effects GLS model using robust standard errors, clustered by participant, and using participant-level controls is used for the analysis:

$$
\begin{aligned}
UserRating_{ij} = \ & b_0 + b_1(High_{ij}) + b_2(InterfaceMatch_i) \\
& + b_3(NumericalDisplay_i) \\
& + b_4(InterfaceMatch_i \times High_{ij}) \\
& + b_5(NumericalDisplay_i \times High_{ij}) \\
& + b_6(ShownRatingNoise_{ij}) \\
& + b_7(PredictedRating_{ij}) \\
& + b_8(AdditionalControls) + u_i + \varepsilon_{ij}
\end{aligned}
$$

$UserRating_{ij}$, $High_{ij}$, $ShownRatingNoise_{ij}$, $PredictedRating_{ij}$, and the *AdditionalControls* are all as defined in Study 1. *InterfaceMatch_i* indicates whether participant i was in a condition in which the recommendation display and response modes matched (i.e., GraphicDisplay-GraphicResponse or NumericDisplay-NumericResponse groups). This variable is used to test for scale compatibility effects. A significant interaction effect *InterfaceMatch_i × High_{ij}* would suggest that scale compatibility affects the degree of bias caused by graphic/numeric similarity between prediction and response formats. *NumericalDisplay_i* indicates whether participant i was in a condition in which the system recommendation was provided numerically rather than graphically (i.e., NumericDisplay-GraphicResponse or NumericDisplay-NumericResponse groups). Testing the interaction *NumericalDisplay_i × High_{ij}* provides a check of the persistence of the graphic/numeric difference between display formats upon bias that was observed in Study 1, after accounting for any scale compatibility effects.

Table 10 shows the regression results. The significant relationship with the *High* variable is consistent with the overall bias observed in Table 9. Of particular interest is the significant interaction effect between High and the interface match indicator (*InterfaceMatch × High*). When there is a match between display and response formats (versus non-match), there is a higher bias. Similarly to what was observed for Study 1, we see about a 40% (i.e., 0.163/(0.163+0.251) based on the values of relevant coefficients from Table 10) bias reduction, here attributable to the mismatch of the interfaces for prediction and response. The analysis also indicates a significant interaction for the numerical display indicator (*NumericalDisplay × High*); however, this effect is not observed with the perturbed ratings and so is less robust (see Appendix B).

Overall, since the bias was not removed with any interface combination, the studies suggest that multiple mechanisms

Table 7. Participant Summary Statistics of Study 2

# of participants (n)	164
Age: Mean (SD)	23.4 (5.68)
Gender	74 M, 90 F
Native speaker of English	68.9% (113/164)
Student level	87 undergrad, 62 grad, 15 others

Table 8. Experimental Design and Sample Sizes per Group in Study 2

Group	N	Example Recommendation Display	Example Rating Response Interface
GraphicDisplay, GraphicResponse	42	Our system thinks that you would rate the joke as: Hate it [____] Love it	Your Rating: Hate it [____] Love it
GraphicDisplay, NumericResponse	40	Our system thinks that you would rate the joke as: Hate it [____] Love it	Your Rating: 3 - Like it ▼
NumericDisplay, GraphicResponse	42	Our system thinks that you would rate the joke as: **3.5 (on a scale of 1 to 5)**	Your Rating: Hate it [____] Love it
NumericDisplay, NumericResponse	40	Our system thinks that you would rate the joke as: **3.5 (on a scale of 1 to 5)**	Your Rating: 3 - Like it ▼

Table 9. Mean User Rating for Each Rating Response and Display Interface Option, After Receiving Artificial Recommendations

Experimental Conditions	High Artificial	Low Artificial	Control
GraphicDisplay, GraphicResponse	3.415	2.800***	3.158
GraphicDisplay, NumericResponse	3.328	2.850***	3.045
NumericDisplay, GraphicResponse	3.371	2.766***	3.233
NumericDisplay, NumericResponse	3.315	2.523***	2.900

***High vs. Low, $p < .001$

impact bias. Still, Study 2 shows statistical support for a scale compatibility mechanism as partially underlying the biases observed in Study 1, as well as the means by which the bias was reduced. Bias is greater when the scales of the predicted rating and the user response match in terms of their graphical/ numerical characteristics. Less clear is whether there is an additional advantage afforded by using a graphical (versus numerical) presentation format. This finding observed in Table 10 is not robust (Appendix B). Still, across all analyses, we observe the least bias when the predicted rating is presented graphically and the user's post-consumption rating is collected numerically.

Discussion and Conclusions

Given the popularity of recommender systems and their increasing importance in consumer decision making, under-

standing the biases generated by such information systems is of significant practical importance. Furthermore, while recommendation biases have been consistently and robustly observed in prior research (Adomavicius et al. 2013; Cosley et al. 2003), recommender systems represent a unique research context that has been underexplored in prior literature on various decision biases and their possible mitigation. In two extensive studies, we tested seven different recommender system display designs for communicating personalized recommendations to the user. Using laboratory experiments and regression analyses controlling for various participant-level factors, we found that none of the seven rating display options—including the canonical ones widely used in practice and the ones representing generic/prototypical features of theoretical interest—completely removed the biases generated by personalized recommendations. However, we do find that some interface displays are more advantageous than others for reducing biases (i.e., the differential biases across various rating display formats represent an important finding that

Table 10. Regression Analysis on Artificial Recommendations for Numeric/Graphic Displays and Responses (Dependent Variable: UserRating)		
	Coefficient Estimate	**Standard Error**
High	0.251†	(0.162)
ShownRatingNoise	0.099	(0.078)
PredictedRating	0.337***	(0.050)
InterfaceMatch	-0.122†	(0.083)
InterfaceMatch×High	0.182*	(0.099)
NumericalDisplay	-0.177*	(0.080)
NumericalDisplay×High	0.163*	(0.100)
Controls		
jokeFunniness	0.654***	(0.106)
Age	-0.008†	(0.005)
male	0.073	(0.058)
undergrad	-0.024	(0.068)
native	-0.093†	(0.063)
IfUsedRecSys	0.075	(0.067)
PredictionAccurate	0.046†	(0.031)
PredictionUseful	0.044*	(0.024)
Constant	-0.688	(0.410)
R^2 within-subject	0.182	
R^2 between-subject	0.564	
R^2 overall	0.274	
X^2	451.70***	

$^\dagger p \leq 0.1$, $*p \leq 0.05$, $**p \leq 0.01$, $***p \leq 0.001$

would not necessarily have been anticipated from prior work). In particular, graphical recommendation formats led to significantly lower biases in users' post-consumption preference ratings than equivalent numerical forms (either as a precise number or a numeric range); however, providing recommendations as ranges of predicted rating values rather than as precise values (either under graphic or numeric conditions) did not impact the appearance of biases. In an additional study, we also varied the user's response format in addition to the recommendation system's display format (using numerical/graphical levels for both factors). The study found consistent evidence for a scale compatibility mechanism at work, such that bias is greater when the recommendation display format is the same as the user response format.

In summary, from the theoretical standpoint, the study extends the classic modify-the-environment paradigm for overcoming decision biases (i.e., by changing the manner in which information is presented to the decision maker) to the highly underexplored issue of biases in *subjective preference construction*. Furthermore, recommender systems provide an additional unique aspect to this study due to the *post-consumption* context of decision biases, that is, where there is

no compelling normative reason for a user to match the system prediction (the user should have all the necessary information to form and report their own subjective preference for the just-consumed item) regardless of the rating format and, hence, where a number of mechanisms traditionally associated with decision biases (e.g., uncertainty or biased recall) would not be operative. And, from the practical standpoint, given the ubiquity of recommender systems in curating content for individuals in a broad variety of application domains, understanding the effects and limitations of such information systems is of immense importance, and the study derives a number of practical insights regarding potential bias mitigation using different rating display designs.

Several key points of theoretical and practical interest follow from the results. First, we reiterate that none of the displays or conditions in the two studies were able to completely eliminate the biasing effects of personalized recommendations on users' reported post-consumption preference ratings for items. However, understanding the more precise theoretical mechanisms behind the two key findings—(1) use of graphical rating display formats and (2) breaking the scale compatibility between the system's rating display format and

user's rating response format—represent important and promising directions for future studies (i.e., toward proposing, developing, and evaluating novel types of interactions and interfaces between recommender systems and their users). However, completely removing recommendation effects is likely a very difficult proposition, which highlights another, broader direction for future IS research: if recommendation biases in post-consumption user ratings cannot be eliminated, the recommender systems' evaluation metrics and procedures (that are largely based on estimating recommendation accuracy based on self-reported user ratings) will need to be rethought.

Another key direction for future research is to explicitly connect the *pre-consumption* use of recommendations to the post-consumption bias of recommendations upon users' stated preferences. Our experiments included only the item consumption and feedback steps, but not the item choice step. However, in real-world systems, consumers typically use the recommendations to select relevant items to consume or purchase. Thaler and Sunstein (2008) argue that general attempts at removing decision biases by modifying the environment should refrain from negatively impacting the options available for choice in the pre-consumption phase described in Figure 1. In the current context, removing decision biases should do so with minimal impact on the usefulness of the personalized recommendations to consumers and firms. In other words, as we reduce the bias created by system recommendations upon preference ratings after-the-fact, we do not want to reduce the usefulness of the recommendations before-the-fact. Conceptually, the two are separable; however, it is an open question as to whether they are separable psychologically. Is bias an unavoidable consequence of providing useful recommendations? This is an important question for future study, particularly given the partial success demonstrated in our research at reducing the bias created by system recommendations.

Future research would also benefit from expanding the scope of the research beyond the personalized recommendations studied here to other forms of recommendation used in practice. For example, the current study focuses on the presentation forms of personalized recommendations generated by systems. In real-world systems, similar interfaces are also commonly used for displaying non-personalized aggregate ratings. Recent research has shown that the aggregate ratings and personalized ratings introduce a similar amount of bias in consumers' preferences (Adomavicius et al. 2016). Hence, we expect the findings of this paper to extend to the aggregate rating context, where the average user ratings serve as non-personalized recommendations, although verification of this is left for future study. Additional possibilities include understanding biases arising from multiple simultaneous rating

displays, as well as understanding the long-term effects of biases on recommender systems.

Another possible direction is for algorithm developers to take on the challenge of developing post-judgment "modifying the person" strategies (Soll et al. 2016), whereby consumers' judgments are debiased *post hoc* by taking into account the system recommendation observed by the user. In other words, *post hoc* debiasing algorithms could try to "reverse-engineer" consumers' true non-biased ratings from the user-submitted ratings and the displayed system recommendations. The challenge is that, although aggregate effects are clear and measurable, it is likely that individual effects (i.e., for a specific user, item, and context) could be variable and irregular. Once the biased ratings are submitted, reverse-engineering can be a difficult task.

From a practical standpoint, our results immediately suggest the potential benefits of changing the practice of using numeric formats (e.g., Star-Numeric, Numeric-Precise) that are currently adopted in many real world applications including Amazon, Yelp, Orbitz, and Expedia for system recommendation displays (i.e., how item ratings, either personalized system-predicted ratings or aggregate user ratings, are shown to the users) and user rating responses (i.e., how users' actual preferences, indicated via post-consumption ratings, are collected by the system). To reduce scale compatibility effects, a preferred practice would be to use different display and response modalities in order to help reduce consumers' biases in their preference ratings. Of course, the insights from our study are limited to the prototypical rating design features that were explored. We expect future studies to propose, develop, and evaluate additional and/or more complex visualizations of rating information, such as combinations of several visualization techniques (e.g., graphical and numeric) in a single rating display or combinations of several different rating displays (e.g., personalized system's ratings and aggregate user ratings), as well as their interactions with different user response modalities.

References

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

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2014. "De-Biasing User Preference Ratings in Recommender Systems," paper presented at the 24th Annual Workshop on Information Technologies and Systems: Value Creation from Innovative Technologies, Auckland, New Zealand.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2016. "Understanding Effects of Personalized Vs. Aggregate Ratings

on User Preferences," in *Proceedings of the Joint Workshop on Interfaces and Human Decision Making for Recommender Systems* (1675), A. Felfernig, P. Brusilovsky, J. O'Donovan, P. Lops, M. C. Willemsen, G. Semeraro, and N. Tintarev (eds.), Aachen, Germany: CEUR-WS, pp. 14-21.

Adomavicius, G., Bockstedt, J., Curley, S., and Zhang, J. 2018. "Effects of Online Recommendations on Consumers' Willingness to Pay," *Information Systems Research* (29:1), pp. 84-102.

Bettman, J. R., and Kakkar, P. 1977. "Effects of Information Presentation Format on Consumer Information Acquisition Strategies," *Journal of Consumer Research* (3:4), pp. 233-240.

Braithwaite, D., and Goldstone, R. 2013. "Flexibility in Data Interpretation: Effects of Representational Format," *Frontiers in Psychology* (4:980), pp. 1-16.

Budescu, D., Por, H.-H., and Broomell, S. 2012. "Effective Communication of Uncertainty in the IPCC Reports," *Climatic Change* (113), pp. 181-200.

Burson, K. A., Larrick, R. P., and Lynch, J. G. 2009. "Six of One, Half Dozen of the Other Expanding and Contracting Numerical Dimensions Produces Preference Reversals," *Psychological Science* (20:9), pp. 1074-1078.

Camilleri, A. R., and Larrick, R. P. 2014. "Metric and Scale Design as Choice Architecture Tools," *Journal of Public Policy & Marketing* (33:1), pp. 108-125.

Casey, B. M., McIntire, D. D., and Leveno, K. J. 2001. "The Continuing Value of the Apgar Score for the Assessment of Newborn Infants," *New England Journal of Medicine* (344), pp. 467-471.

Cosley, D., Lam, S., Albert, I., Konstan, J. A., and Riedl, J. 2003. "Is Seeing Believing? How Recommender Interfaces Affect Users' Opinions," in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, New York: ACM, pp. 585-592.

Dieckmann, N., Mauro, R., and Slovic, P. 2010. "The Effects of Presenting Imprecise Probabilities in Intelligence Forecasts," *Risk Analysis* (30:6), pp. 987-1001.

Du, N., Budescu, D., Shelly, M., and Omer, T. 2011. "The Appeal of Vague Financial Forecasts," *Organizational Behavior and Human Decision Processes* (114:2), pp. 179-189.

Fischer, G. W., and Hawkins, S. A. 1993. "Strategy Compatibility, Scale Compatibility, and the Prominence Effect," *Journal of Experimental Psychology: Human Perception and Performance* (19:3), pp. 580-597.

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

Godinho de Matos, M., Ferreira, P., Smith, M. D., and Telang, R. 2016. "Culling the Herd: Using Real-World Randomized Experiments to Measure Social Bias with Known Costly Goods," *Management Science* (62:9), pp. 2563-2580.

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

Gomez-Uribe, C. A., and Hunt, N. 2016. "The Netflix Recommender System: Algorithms, Business Value, and Innovation," *ACM Transactions on Management Information Systems* (6:4:13).

Haselton, M. G., Nettle, D., and Andrews, P. W. 2015. "The Evolution of Cognitive Bias," in *The Handbook of Evolutionary Psychology*, D. M. Buss (ed.), Hoboken, NJ: John Wiley & Sons, Inc., pp. 724-746.

Hirt, E., Kardes, F., and Markman, K. 2004. "Activating a Mental Simulation Mind-Set through Generation of Alternatives: Implications for Debiasing in Related and Unrelated Domains," *Journal of Experimental Social Psychology* (40:3), pp. 374-383.

Hsee, C. K. 1995. "Elastic Justification: How Tempting but Task-Irrelevant Factors Influence Decisions," *Organizational Behavior and Human Decision Processes* (62:3), pp. 330-337.

Hu, N., Pavlou, P. A., and Zhang, J. 2017. "On Self-Selection Biases in Online Product Reviews," *MIS Quarterly* (41:2), pp. 449-471.

Jarvenpaa, S. L. 1989. "The Effect of Task Demands and Graphical Format on Information Processing Strategies," *Management Science* (35:3), pp. 285-303.

Joslyn, S., and LeClerc, J. 2012. "Uncertainty Forecasts Improve Weather-Related Decisions and Attenuate the Effects of Forecast Error," *Journal of Experimental Psychology* (18:1), pp. 126-140.

Kahneman, D., Slovic, P., and Tversky, A. (eds.). 1982. *Judgment under Uncertainty: Heuristics and Biases*, Oxford, UK: Cambridge University Press.

Keeney, R. L. 2012. "Value-Focused Brainstorming," *Decision Analysis* (9), pp. 303-313.

Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. 1997. "Grouplens: Applying Collaborative Filtering to Usenet News," *Communications of the ACM* (40:3), pp. 77-87.

Larrick, R. 2004. "Debiasing," in *Blackwell Handbook of Judgment and Decision Making*, D. J. Koehler and N. Harvey (eds.), Malden, MA: Blackwell.

Larrick, R. P., and Soll, J. B. 2008. "The MPG Illusion," *Science* (320:5883), pp. 1593-1594.

Li, X., and Hitt, L. M. 2008. "Self-Selection and Information Role of Online Product Reviews," *Information Systems Research* (19:4), pp. 456-474.

Lu, J., Wu, D., Mao, M., Wang, W., and Zhang, G. 2015. "Recommender System Application Developments: A Survey," *Decision Support Systems* (74), pp. 12-32.

Maniadis, Z., Tufano, F., and List, J. A. 2014. "One Swallow Doesn't Make a Summer: New Evidence on Anchoring Effects," *The American Economic Review* (104:1), pp. 277-290.

Meyer, J., Shamo, M., and Gopher, D. 1999. "Information Structure and the Relative Efficacy of Tables and Graphs," *Human Factors* (41:4), pp. 570-587.

Porat, T., Oron-Gilad, T., and Meyer, J. 2009. "Task-Dependent Processing of Tables and Graphs," *Behaviour & Information Technology* (28:3), pp. 293-307.

Ricci, F., Rokach, L., and Shapira, B. 2015. *Recommender Systems Handbook*, New York: Springer.

Russo, J. E., and Schoemaker, P. J. H. 2014. "Overconfidence," in *The Palgrave Encyclopedia of Strategic Management,* M. Augier and D. Teece (eds.), London: Palgrave Macmillan.

Salganik, M. J., and Watts, D. J. 2008. "Leading the Herd Astray: An Experimental Study of Self-Fulfilling Prophecies in an Artificial Cultural Market," *Social Psychology Quarterly* (74:4), pp. 338-355.

Sarwar, B., Karypis, G., Konstan, J. A., and Riedl, J. 2001. "Item-Based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International Conference on World-Wide-Web*, New York: ACM, pp. 285-295.

Schkade, D. A., and Johnson, E. J. 1989. "Cognitive Processes in Preference Reversals," *Organizational Behavior and Human Decision Processes* (44:2), pp. 203-231.

Sharma, A., Hofman, J. M., and Watts, D. J. 2015. "Estimating the Causal Impact of Recommendation Systems from Observational Data," in *Proceedings of the 16th ACM Conference on Economics and Computation*, New York: ACM, pp. 453-470.

Simonson, I., and Drolet, A. 2004. "Anchoring Effects on Consumers' Willingness-to-Pay and Willingness-to-Accept," *Journal of Consumer Research* (31:3), pp. 681-690.

Slovic, P., Griffin, D., and Tversky, A. 1990. "Compatibility Effects in Judgment and Choice," in *Insights in Decision Making: A Tribute to Hillel J. Einhorn*, R. M. Hogarth (ed.), Chicago: University of Chicago Press, pp. 5-27.

Soll, J. B., Milkman, K. L., and Payne, J. W. 2016. "A User's Guide to Debiasing," in *The Wiley Blackwell Handbook of Judgment and Decision Making*, Hoboken, NJ: John Wiley & Sons, Ltd, pp. 924-951.

Thaler, R., and Sunstein, C. 2008. *Nudge: Improving Decisions About Health, Wealth, and Happiness*, New Haven CT: Yale University Press.

Thorsteinson, T., Breier, J., Atwell, A., Hamilton, C., and Privette, M. 2008. "Anchoring Effects on Performance Judgments," *Organizational Behavior and Human Decision Processes* (107:1), pp. 29-40.

Tufano, F. 2010. "Are 'True' Preferences Revealed in Repeated Markets? An Experimental Demonstration of Context-Dependent Valuations," *Experimental Economics* (13:1), pp. 1-13.

Tversky, A., and Kahneman, D. 1974. "Judgment under Uncertainty: Heuristics and Biases," *Science* (185:4157), pp. 1124-1131.

Tversky, A., Sattath, S., and Slovic, P. 1988. "Contingent Weighting in Judgement and Choice," *Psychological Review* (95:3), pp. 371-384.

Vessey, I. 1991. "Cognitive Fit: A Theory-Based Analysis of the Graphs Versus Tables Literature," *Decision Sciences* (22:2), pp. 219-240.

About the Authors

Gediminas Adomavicius is a professor of Information and Decision Sciences at the University of Minnesota, where he also holds Carolyn I. Anderson Chair in Business Education Excellence. He received his Ph.D. in Computer Science from New York University. His research interests include recommender systems, machine learning, and electronic market mechanisms. His work has been published in leading Information Systems and Computer Science journals. He received the NSF CAREER Award for his research on recommender systems, has served as senior editor for *Information Systems Research* and *MIS Quarterly*, and is a Distinguished Fellow of the INFORMS Information Systems Society.

Jesse Bockstedt is an associate professor of Information Systems and Operations Management in the Goizueta Business School at Emory University. He received his Ph.D. in business administration with a focus on Information Systems from the University of Minnesota. He studies user behavior and economic issues in environments that rely on information technology. His research has appeared in a variety of journals, including *Information Systems Research, MIS Quarterly, Journal of MIS, Journal of Operations Management,* and *Production and Operations Management*.

Shawn P. Curley is a professor of Information and Decision Sciences at the Carlson School of Management, University of Minnesota. He received his Ph.D. in Psychology from the University of Michigan. His research interests are in behavioral decision theory and include user behavior with recommender systems and user acceptance of complex auction mechanisms. Recent research outlets include *Information Systems Research, Management Science, MIS Quarterly,* and *Journal of Operations Management*.

Jingjing Zhang is an associate professor of Operations and Decision Technologies at the Kelley School of Business, Indiana University. She received her Ph.D. in business administration with an Information Systems concentration from the University of Minnesota. Her research interests include personalization techniques, recommender systems, and human–computer interactions. Her research has appeared in journals including *Information Systems Research, INFORMS Journal on Computing, IEEE Transactions on Knowledge and Data Engineering,* and *ACM Transactions on Information Systems*.

REDUCING RECOMMENDER SYSTEM BIASES: AN INVESTIGATION OF RATING DISPLAY DESIGNS

Gediminas Adomavicius
Information and Decision Sciences, Carlson School of Management, University of Minnesota,
321 19th Avenue South, Minneapolis, MN 55455 U.S.A. {gedas@umn.edu}

Jesse C. Bockstedt
Information Systems and Operations Management, Goizueta Business School, Emory University,
1300 Clifton Road, Atlanta, GA 30322 U.S.A. {bockstedt@emory.edu}

Shawn P. Curley
Information and Decision Sciences, Carlson School of Management, University of Minnesota,
321 19th Avenue South, Minneapolis, MN 55455 U.S.A. {curley@umn.edu}

Jingjing Zhang
Operations and Decision Technologies, Kelley School of Business, Indiana University,
1309 East Tenth Street, Bloomington, IN 47405 U.S.A. {jjzhang@indiana.edu}

Appendix A

Study 1 Experimental Results: Perturbed Recommendations in Various Display Types

As an extension to a more realistic setting and as a robustness check, we examine whether biases generated by *perturbations* in real recommendations from an actual recommender system can be eliminated by the rating display options. Recall that participants received some recommendations that were perturbed either upward (High-Perturbed) or downward (Low- Perturbed) by 1 star from the actual predicted ratings. As a control, each participant also received recommendations without perturbations (Accurate).

For our main analysis of the perturbed recommendations, submitted ratings for the jokes were adjusted for the predicted ratings in order to obtain a response variable on a comparable scale across subjects. Thus, the main response variable is the *rating drift*, which is defined by Adomavicius et al. (2013) as[1]:

$$RatingDrift = UserRating - PredictedRating$$

The *RatingDrift* variable captures the degree to which the user's submitted rating is higher or lower than the true rating predicted by the system (i.e., *before* any experiment-specific perturbations are applied to it), thereby accounting for individual differences in preference for different

---

[1]Following Adomavicius et al., we use rating drift as the dependent variable for analyzing bias effects with perturbed recommendations. Since *PredictedRating* is a component of rating drift and also included as a control factor in the model, we could alternatively run the model using *UserRating* as the dependent variable. Since this analysis is interested in the effects of perturbations from the predicted rating, we follow Adomavicius et al. in using rating drift (also measuring differences from predicted ratings) as a more natural and intuitive measure in this case. In any event, all conclusions are identical using *UserRating* as the dependent variable (only the *PredictedRating* coefficient estimates change in the model).

items being viewed. Thus, using *RatingDrift* puts our dependent measure on an equal footing (across different users and items) with the perturbations that are the main manipulation for these conditions.

As done with the artificial ratings, we first test for particular differences among the seven different rating display interfaces. Figure A1 is a plot of the aggregate means of rating drift for each treatment group when recommendations were perturbed to be higher or lower or received no perturbation. As can be seen, the negative perturbations (Low, triangle) led to negative rating drifts and positive perturbations (High, dot) led to positive drifts in user ratings, while the accurate recommendations with no perturbation (Accurate, square) led to drifts around zero. For each rating display, we performed pair-wise *t*-tests to compare user-submitted ratings after receiving high and low artificial recommendations. The *t*-test results are presented in Table A1.
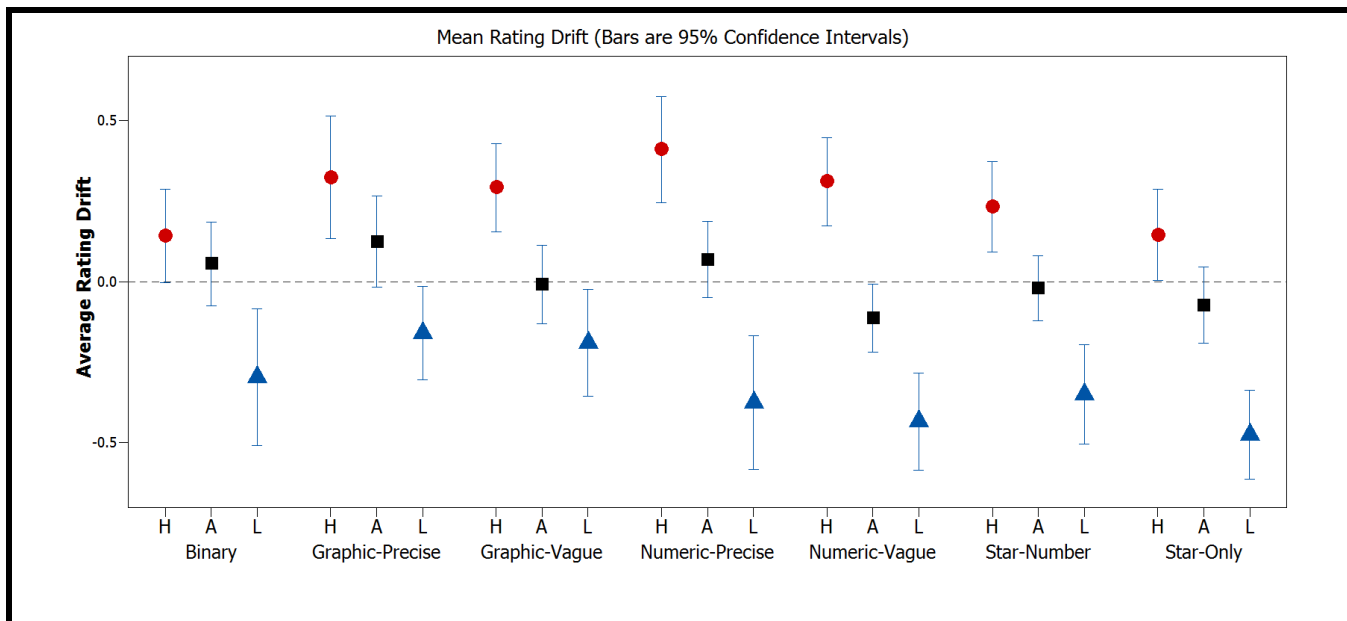

Figure A1. Means and 95% Confidence Intervals of User Rating Drift After Receiving High Perturbed (High: dot), Low Perturbed (Low: triangle), and Non-Perturbed Recommendations (Accurate: square)

Table A1. Pair-Wise Comparisons of Mean Rating Drift Difference for Each Rating Display Option Using *T*-Tests

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

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

All mean rating drift comparisons between High and Low perturbed conditions are significant for all rating display options (one-tailed *p*-value < 0.001 for all High versus Low tests), showing a clear and positive bias of system recommendations on consumers' rating drift. Hence, similarly to the artificial recommendation scenario, we found that none of the seven rating display options could completely remove the biases generated by perturbed real recommendations.

We next performed regression analyses to compare the size of recommendation bias across different rating display options, while controlling for participant-level factors. The random effects GLS model using robust standard errors, clustered by participant, and participant-level controls represents our model for the analysis:

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

We ran the model using the Numeric-Precise as the baseline, paralleling the analysis in Table 5 of the paper. Table A2 summarizes the regression analysis of perturbed recommendations.

Similar to what we found in the artificial recommendation analysis, several of the interaction terms were significant. Although the differences that attain statistical significance at a 5% level are fewer in number, they correspond to those seen in Table 5 with the artificial recommendations. The Numeric-Precise display yielded the greatest system recommendation effect, and the effect was significantly greater than for the Graphic-Vague and Binary displays. In addition, the Binary display showed the least effect, and the effect was significantly lower than for the Numeric-Precise and Numeric-Vague displays.

Table A2. Regression Analysis on High and Low Perturbed Recommendations (Baseline: Numeric-Precise; Dependent Variable: RatingDrift)		
	Coefficient Estimate	**Standard Error**
High	0.778***	(0.119)
PredictedRating	-0.123*	(0.068)
Display		
Numeric-Vague	-0.083	(0.127)
Star-Numeric	0.008	(0.131)
Star-Only	-0.126	(0.126)
Graphic-Precise	0.198	(0.126)
Graphic-Vague	0.163	(0.132)
Binary	0.065	(0.143)
Interactions		
Numeric-Vague×High	-0.040	(0.153)
Star-Numeric×High	-0.189	(0.157)
Star-Only×High	-0.140	(0.154)
Graphic-Precise×High	-0.285	(0.168)
Graphic-Vague×High	-0.302*	(0.152)
Binary×High	-0.360*	(0.169)
Controls		
jokeFunniness	0.234*	(0.095)
age	0.000	(0.006)
male	0.029	(0.045)
undergrad	-0.051	(0.067)
native	-0.015	(0.056)
IfUsedRecSys	0.049	(0.059)
PredictionAccurate	0.038	(0.031)
PredictionUseful	0.009	(0.025)
Constant	-0.988**	(0.386)
R^2 within-subject	0.149	
R^2 between-subject	0.010	
R^2 overall	0.121	
x^2	266.15***	

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

Paralleling the analysis in Table 6 of the paper, for our perturbed recommendations we next conduct the regression analysis for our baseline 2×2 between-subjects design on the two dimensions of display format (numeric versus graphic) and precision of the recommendations (precise versus vague). We created a panel from the data as each participant was exposed to both high and low perturbed recommendations in a random fashion. The regression model used generalized least squares (GLS) estimation, a random effect to control for participant-level heterogeneity, and robust standard errors clustered by participant:

$$RatingDrift_{ij} = b_0 + b_1(High_{ij}) + b_2(Presentation_i) + b_3(Precision_i) +$$
$$b_4(Presentation_i \times Precision_i) + b_5(Presentation_i \times High_{ij}) + b_6(Precision_i \times High_{ij}) +$$
$$b_7(ShownRatingNoise_{ij}) + b_8(PredictedRating_{ij}) + b_9(AdditionalControls) + u_i + \varepsilon_{ij}$$

$RatingDrift_{ij}$ is the difference between the submitted rating and the predicted rating for participant i on joke j. The right-hand-side variables are the same as those used in the analysis of the artificial recommendation data in the main body of the paper.

| Table A3. Regression Analysis on Perturbed Recommendations, for Numeric/Graphic and Precise/ Vague Rating Displays (Dependent Variable: RatingDrift)		
	Coefficient Estimate	**Standard Error**
High	0.469***	(0.086)
PredictedRating	-0.203**	(0.083)
Presentation (Numeric = 1, Graphic = 0)	-0.254**	(0.098)
Precision (Precise = 1, Vague = 0)	0.034	(0.094)
Presentation×Precision	0.081	(0.118)
Presentation×High	0.267**	(0.108)
Precision×High	0.028	(0.107)
Controls		
jokeFunniness	0.413***	(0.129)
age	0.003	(0.007)
male	0.099	(0.063)
Undergrad	-0.137	(0.089)
native	-0.069	(0.066)
IfUsedRecSys	0.096	(0.075)
PredictionAccurate	0.087*	(0.036)
PredictionUseful	-0.040	(0.027)
Constant	-1.263*	(0.503)
R^2 within-subject	0.174	
R^2 between-subject	0.041	
R^2 overall	0.147	
χ^2	184.15***	

*$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

Our results (Table A3) corroborate the findings from the analyses using artificial recommendations. Specifically, the results confirm that the precision of the recommendations does not have a significant main effect ($p = 0.372$) or interaction with rating perturbation (i.e., *Precision × High*, $p = 0.390$) in affecting rating shifts. As shown in Figure A2b, recommendations displayed in precise versus vague forms have a similar level of effects in influencing rating drift (slopes are not detectably different).
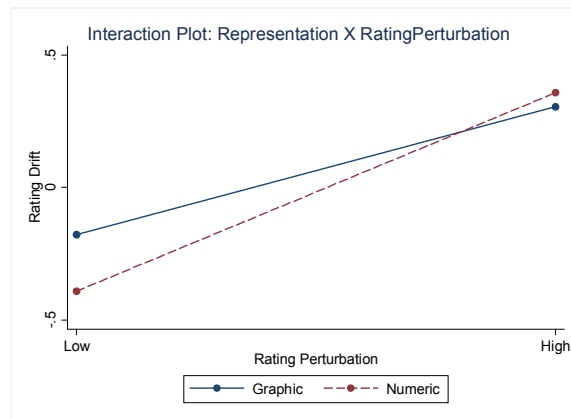
(a) **Information Representation (Numeric or Graphic) and Perturbed Rating (High or Low)**
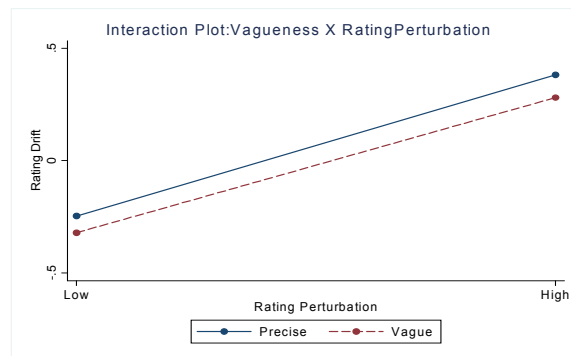
(b) **Precision of Recommendation (Precise or Vague) and Perturbed Rating (High or Low)**

Note: The dotted reference line denotes the mean user pre-treatment ratings.

Figure A2. Interaction Plots of Rating Displays Forms and Rating Values

However, presenting recommendations in numeric format compared to a graphical format exhibits a significant main effect ($p = 0.005$) and interaction (i.e., *Presentation × High*, $p = 0.0075$) upon rating drift. The interaction plot in Figure A2a illustrates that the combined effect—the difference between ratings for High versus Low conditions (slope of the line)—is greater in the numeric conditions compared to the graphical conditions.

Appendix B

Study 2 Experimental Results: Perturbed Recommendations in Various Display and Response Types

Following the analyses of Study 1, we compared the mean rating drift of the four experimental groups after receiving perturbed recommendations. Table B1 summarizes the aggregate means for each of the experimental groups and within-subjects conditions. All comparisons between High and Low conditions are significant across the four experimental conditions, showing that a significant bias exists for all display-response combinations. As with Study 1, none of the interface designs were able to remove the bias completely.

Table B1. Mean Rating Drift for Each Rating Response and Display Interface Option, After Receiving Perturbed Recommendations			
Experimental Conditions	**High Perturbed**	**Low Perturbed**	**Accurate**
GraphicDisplay, GraphicResponse	0.443	-0.260***	0.119
GraphicDisplay, NumericResponse	0.331	-0.160***	0.125
NumericDisplay, GraphicResponse	0.282	-0.271***	-0.035
NumericDisplay, NumericResponse	0.412	-0.375***	0.070

*** High versus Low, $p < .001$

For the system recommendations that are perturbed to be either higher or lower, we again use rating drift as a measure of bias and a model paralleling that for Study 1:

$$RatingDrift_{ij} = b_0 + b_1(High_{ij}) + b_2(InterfaceMatch_i) + b_3(NumericalDisplay_i) + b_4(InterfaceMatch_i \times High_{ij}) + b_5(NumericalDisplay_i \times High_{ij}) + b_6(PredictedRating_{ij}) + b_7(AdditionalControls) + u_i + \varepsilon_{ij}$$

All variable definitions are consistent with those described in the paper. Table B2 shows the regression results. The results somewhat differ for the perturbed recommendations. Only the matching display-response formats variable shows a higher bias for perturbed recommendations. The numerical versus graphical display manipulation did not influence participants' bias.

Table B2. Regression Analysis on Perturbed Recommendations for Numeric/Graphic Displays and Responses (Dependent Variable: RatingDrift)	Coefficient Estimate	Standard Error
High	0.497***	(0.100)
PredictedRating	-0.046	(0.085)
InterfaceMatch	-0.079	(0.084)
InterfaceMatch×High	0.220*	(0.111)
NumericalDisplay	-0.094	(0.082)
NumericalDisplay×High	0.059	(0.110)
Controls		
jokeFunniness	0.367***	(0.120)
Age	-0.002	(0.006)
Male	0.071	(0.065)
Undergrad	-0.073	(0.083)
Native	-0.029	(0.062)
IfUsedRecSys	0.115†	(0.078)
PredictionAccurate	0.013	(0.037)
PredictionUseful	0.014	(0.032)
Constant	-1.378	(0.477)
R^2 within-subject	0.154	
R^2 between-subject	0.035	
R^2 overall	0.129	
X^2	175.20***	

†$p \leq 0.1$, *$p \leq 0.05$, **$p \leq 0.01$, ***$p \leq 0.001$

Overall, Study 2 shows consistently significant support for the effect of scale compatibility upon bias. However, the graphical/numerical processing differences for incoming information describe a less stable effect.

Appendix C

Performance of Recommendation Algorithms ▮▮▮▮▮▮

We used the well-known item-based collaborative filtering (CF) technique (Sarwar et al. 2001) to implement a recommender system that estimated users' preference ratings for the jokes. We chose to use this technique in our experiment for several reasons. First, item-based CF is one of the most popular techniques used in real-world applications because of its efficiency and accuracy (e.g., Adomavicius and Tuzhilin 2005; Deshpande and Karypis 2004; Linden et al. 2003; Sarwar et al. 2001). Second, this technique allows us to precompute the main portion of our recommendation model (i.e., the similarity scores between items based on their rating patterns) *in advance* based on the extensive Jester rating dataset. Thus, we could compute rating predictions for our experimental participants as soon as they submitted their preferences. We did not have to perform extensive recomputations for each new participant on the fly (as needed for some other recommendation techniques), which was important due to the real-time nature of our experiment. Third, while an item-based approach was very computationally efficient, its accuracy performance was also either better or not significantly different than a number of other techniques that we explored. The details of this selection process are provided below.

Table C1. Comparison of Recommendation Algorithms on Joke Rating Dataset		
Algorithm	**Description**	**Predictive Accuracy (MAE)**
User Average	Predict each unknown user-item rating as an average of all ratings of that user (item)	0.7458
Item Average		0.7686
Baseline Estimate	Computes each unknown user-item rating with the baseline estimate which is a combination of the global mean of ratings in the dataset, the average rating deviation of corresponding user and the average rating deviation of the corresponding item.	0.6897
Weighted Slope One	Estimates average rating difference between all item pairs. For a given unknown user-item rating, finds all the items that were co-rated with this item and computes predictions based on each of these items. The final prediction is a weighted sum of predictions from individual items (Lemire and Maclachlan 2005).	0.6897
User-based CF	For each unknown rating, finds the most similar items that have been rated by the same user (or the most similar users who have rated the same item) and predicts the rating as a weighted sum of neighbors' ratings (Breese et al. 1998; Sarwar et al. 2001).	0.6841
Item-based CF		**0.6795**
Matrix Factorization	Decomposes the rating matrix into two matrices so that each user and each item is associated with a user vector and an item vector of latent variables. Prediction is done by taking inner product of user and item vectors (Funk 2006; Koren et al. 2009).	0.6817

Based on the joke rating data collected in Task 1, we built a recommender system to predict user's preference ratings on jokes. We compared seven popular recommendation techniques (Table C1) to find the best-performing technique for our data set. The techniques included simple user- and item-based rating average methods, user-item baseline method (Bell and Koren 2007), weighted Slope One approach (Lemire and Maclachlan 2005), user- and item-based collaborative filtering (CF) approaches (Breese et al. 1998; Sarwar et al. 2001), as well as a model-based matrix factorization algorithm (Funk 2006; Koren et al. 2009). Each recommendation algorithm was evaluated using five-fold cross validation based on the standard Mean Absolute Error (MAE) metric. As we can see from the Table C1, the item-based CF algorithm had the best MAE performance (0.6795), followed by matrix factorization (0.6817), then by user-based CF (0.6841), weighted slope one (0.6897) and baseline (0. 6897), and lastly by the simple user average (0.7458) and item average (0.7686) approaches. Based on these results, we selected the item-based CF approach as our recommendation technique to predict users' preference ratings on jokes.

To further validate the performance of the selected item-based CF algorithm, we applied the algorithm on the jokes that were included in the control conditions. These control jokes were rated by users as part of the experiment but did not have any recommendation displayed. We then compared the predicted rating against the actual user ratings, and the mean difference, calculated as an average of pair-wise differences between corresponding actual and predicted ratings, is 0.005. Our pair-wise *t*-test results show that the differences between the system-predicted and actual ratings are not significant (two-tailed *p*-value = 0.8202). We further conducted the comparison test separately for each treatment group and found that none of the groups had significant differences between system-predicted ratings and actual user ratings. Table C2 summarizes our results.
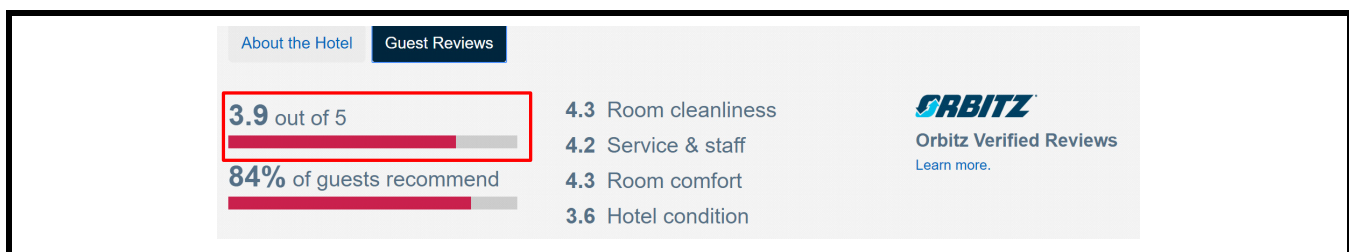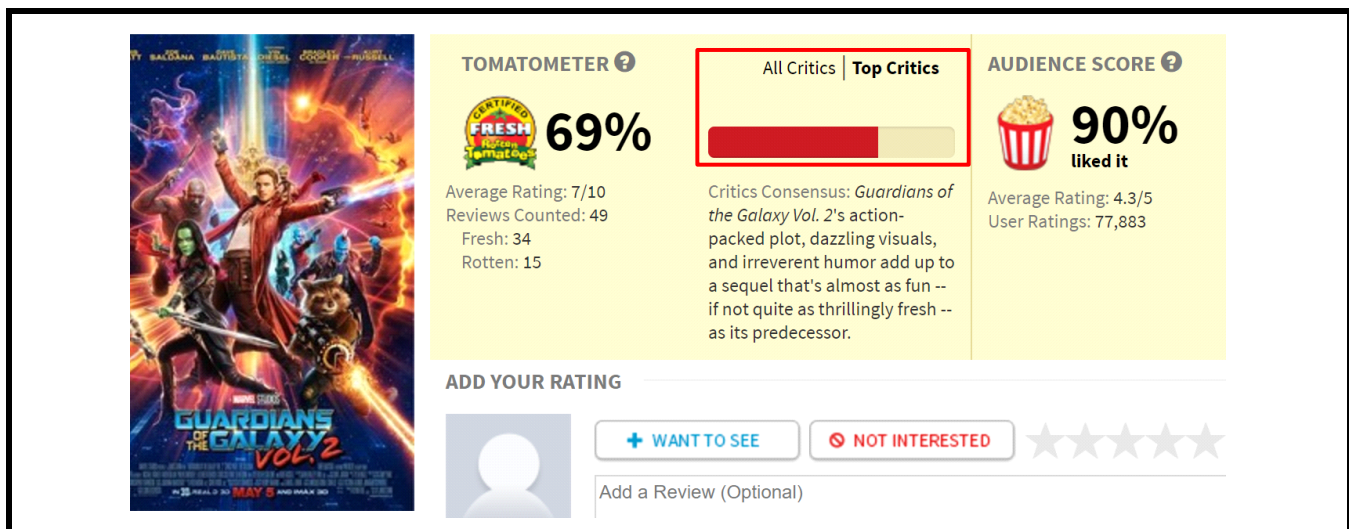
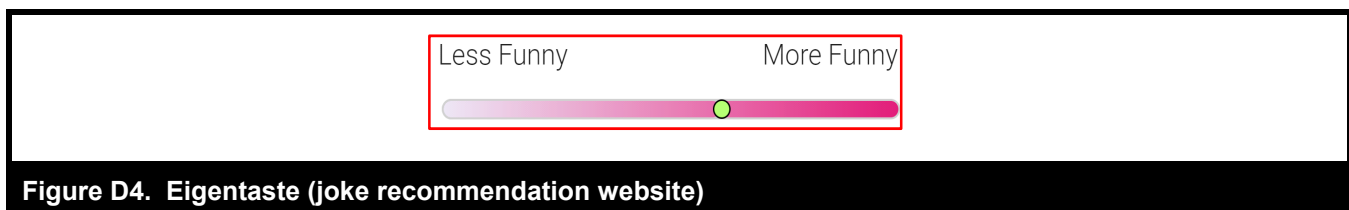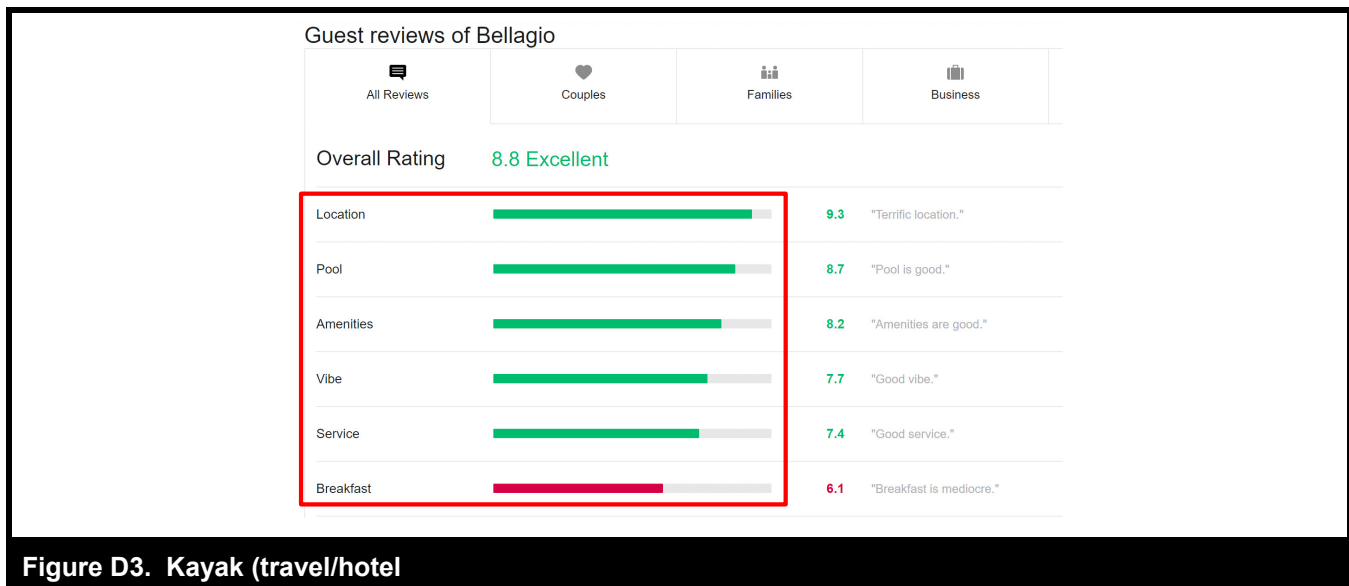| Table C2. Mean Difference between Predicted and Actual Ratings on Control Jokes				
Group	Mean Predicted Rating	Mean User Rating	Mean Difference	*P*-value of pair-wise t-test
Binary	2.8495	2.8900	0.0405	0.5162
Graphic-Precise	3.0175	3.0450	0.0275	0.6917
Graphic-Vague	2.8575	2.8650	0.0075	0.9179
Numeric-Precise	2.9790	2.9000	-0.0790	0.2313
Numeric-Vague	2.8144	2.8103	-0.0041	0.9551
Star-Number	2.8947	2.9311	0.0364	0.5156
Star-Only	2.8772	2.8837	0.0065	0.9171

Appendix D

Examples of Graphical Rating Bar Displays

Below are screenshots of different graphical rating bar examples (e.g., accompanied by numeric values or not, shaped like a progress bar versus a slider bar, etc.) taken from different websites. Red borders were added for emphasis.

Figure D1. Orbitz (travel/hotel)

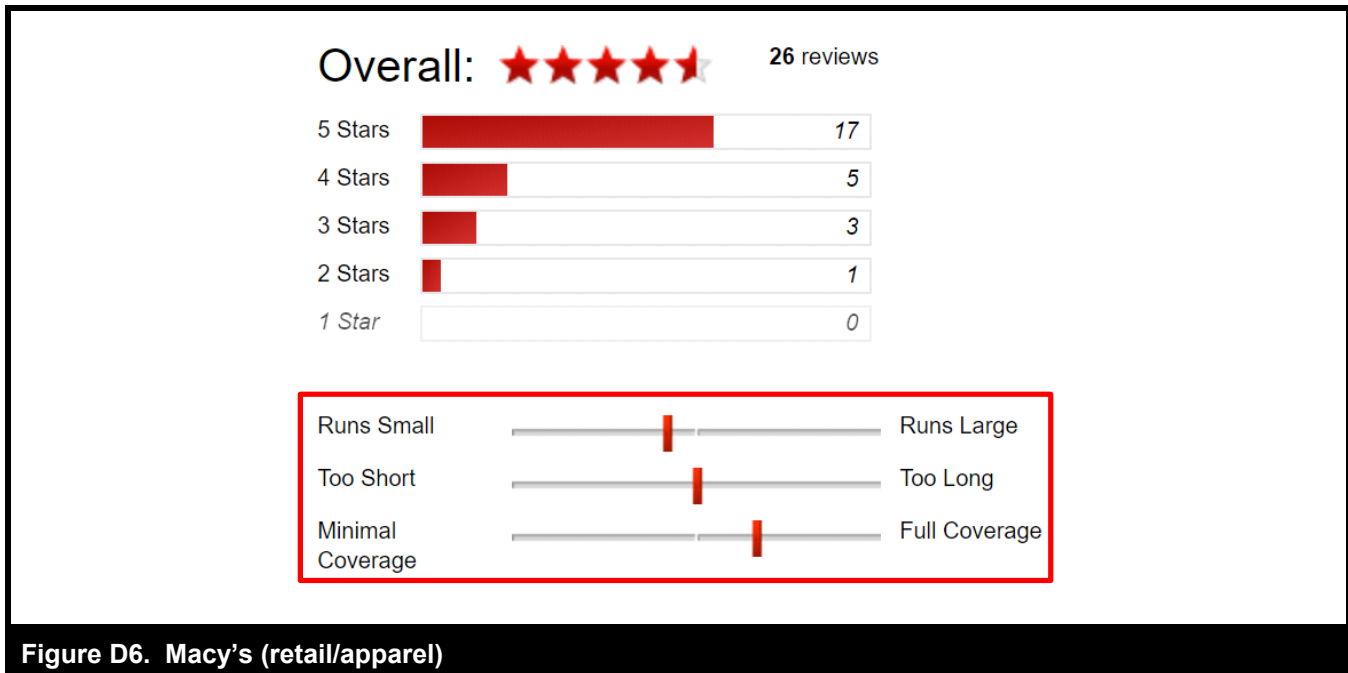**Figure D2. Rotten Tomatoes (movies)**

Guest reviews of Bellagio

All Reviews	Couples	Families	Business

Overall Rating   8.8 Excellent

Location		9.3	"Terrific location."
Pool		8.7	"Pool is good."
Amenities		8.2	"Amenities are good."
Vibe		7.7	"Good vibe."
Service		7.4	"Good service."
Breakfast		6.1	"Breakfast is mediocre."

Figure D3. Kayak (travel/hotel

Less Funny   More Funny

Figure D4. Eigentaste (joke recommendation website)

Average Customer Ratings

Overall   ★★★★☆   3.8

Overall size		
Too big		Too small

Length		
Too long		Too short

Waist		
Too loose		Too tight

Chest		
Too loose		Too tight

Sleeves		
Too long		Too short

Figure D5. Banana Republic (retail/apparel)

Figure D6. Macy's (retail/apparel)

Examples of System Ratings and User Rating Interfaces Displayed Together

Below are screenshots of several online services that commonly display system ratings next to the interface for users to submit their personal ratings for items.
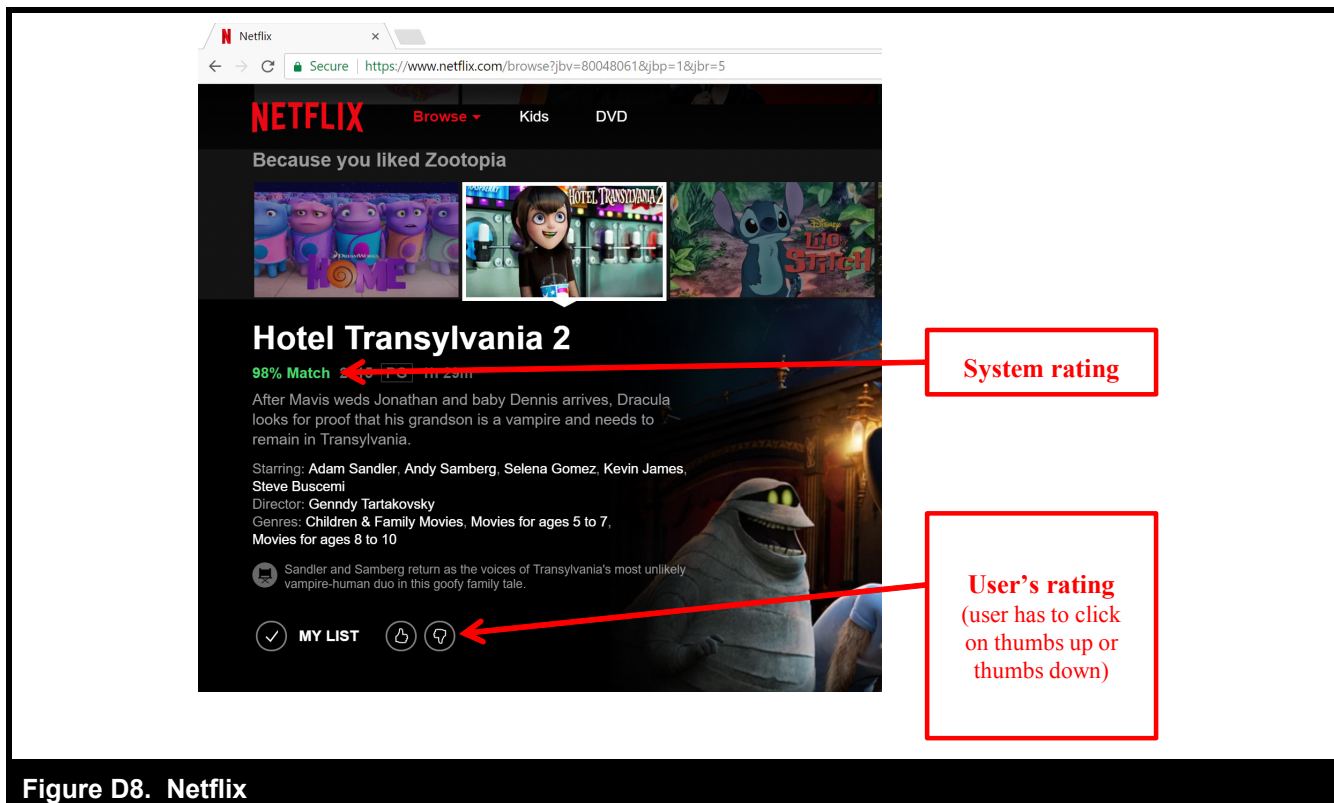

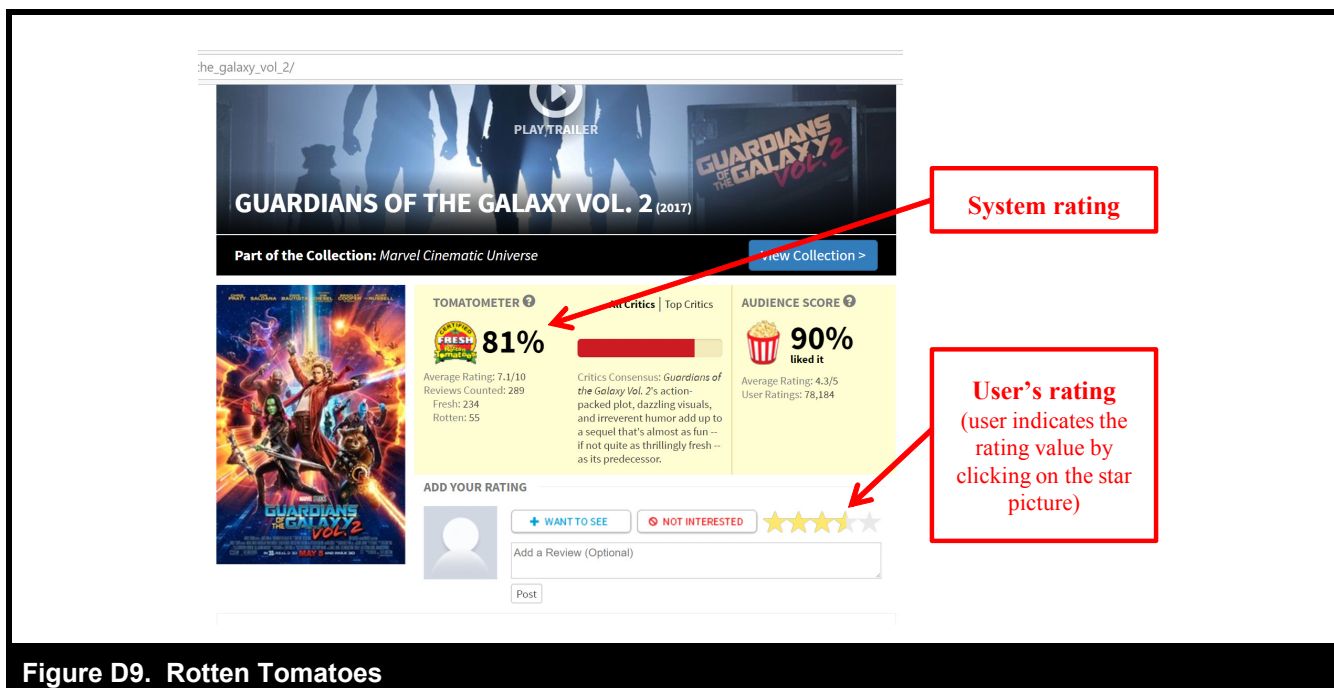
Figure D7. IMDB

Figure D8. Netflix

Figure D9. Rotten Tomatoes

References

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

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

Bell, R. M., and Koren, Y. 2007. "Improved Neighborhood-Based Collaborative Filtering," in *Proceedings of the 13th ACM SIGKDD International Conference on Knowledge (KDD Cup'07)*, New York: ACM, pp. 7-14.

Breese, J. S., Heckerman, D., and Kadie, C. 1998. "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," in *Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence*, San Francisco: Morgan Kaufmann, pp. 43-52.

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

Funk, S. 2006. "Netflix Update: Try This at Home" (http://sifter.org/~simon/journal/20061211.html).

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

Lemire, D., and Maclachlan, A. 2005. "Slope One Predictors for Online Rating-Based Collaborative Filtering," in *Proceedings of the 2005 SIAM International Conference on Data Mining,* H. Kargupta, J. Srivastava, C. Kamath and A. Goodman (eds.), Newport Beach, California, pp. 471-475.

Linden, G., Smith, B., and York, J. 2003. "Amazon.Com Recommendations: Item-to-Item Collaborative Filtering," *IEEE Internet Computing* (7:1), pp. 76-80.

Sarwar, B., Karypis, G., Konstan, J. A., and Riedl, J. 2001. "Item-Based Collaborative Filtering Recommendation Algorithms," in *Proceedings of the 10th International World Wide Web Conference*, New York: ACM, pp. 285-295.