Information Systems Research

Effects of Online Recommendations on Consumers' Willingness to Pay

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

Please scroll down for article—it is on subsequent pages

Effects of Online Recommendations on Consumers' Willingness to Pay

Gediminas Adomavicius,[a] Jesse C. Bockstedt,[b] Shawn P. Curley,[a] Jingjing Zhang[c]

[a] Information and Decision Sciences, Carlson School of Management, University of Minnesota, Minneapolis, Minnesota 55455;
[b] Information Systems and Operations Management, Goizueta Business School, Emory University, Atlanta, Georgia 30322;
[c] Operations and Decision Technologies, Kelley School of Business, Indiana University, Bloomington, Indiana 47405
Contact: gedas@umn.edu (GA); bockstedt@emory.edu, http://orcid.org/0000-0002-4274-9744 (JCB); curley@umn.edu (SPC); jjzhang@indiana.edu, http://orcid.org/0000-0002-6805-8685 (JZ)

Abstract. Recommender systems are an integral part of the online retail environment. Prior research has focused largely on computational approaches to improving recommendation accuracy, and only recently researchers have started to study their behavioral implications and potential side effects. We used three controlled experiments, in the context of purchasing digital songs, to explore the willingness-to-pay judgments of individual consumers after being shown personalized recommendations. In Study 1, we found strong evidence that randomly assigned song recommendations affected participants' willingness to pay, even when controlling for participants' preferences and demographics. In Study 2, participants viewed actual system-generated recommendations that were intentionally perturbed (introducing recommendation error), and we observed similar effects. In Study 3, we showed that the influence of personalized recommendations on willingness-to-pay judgments was obtained even when preference uncertainty was reduced through immediate and mandatory song sampling prior to pricing. The results demonstrate the existence of important economic side effects of personalized recommender systems and inform our understanding of how system recommendations can influence our everyday preference judgments. The findings have significant implications for the design and application of recommender systems as well as for online retail practices.

1. Introduction

Recommender systems assist consumers in finding interesting items from a large set of possible items. One of the key objectives of many recommender systems is to predict personal preferences of an individual consumer[1]—often expressed as numeric ratings—for items that they have not yet purchased, experienced, or considered (depending on the specific application context). For example, Netflix, the Internet television and movie streaming/rental company, has used 1–5 star-rating predictions for different movies to its users, and Yahoo! Music used a similar scale for songs. Some online dating sites, such as OkCupid, also use numeric compatibility ratings as recommendations. Recommender systems have become commonplace in online purchasing environments, where they help to reduce search costs and protect people from being overwhelmed when the number of products available for purchase is large. Personalized recommendations[2] not only add value to users but can also benefit sellers. For example, Amazon has reported that 35% of its product sales

result from recommendations (Marshall 2006). Netflix reported that about 75% of the content watched by its subscribers was suggested by its recommendation system (Amatriain and Basilico 2012). Research in information systems and computer science has focused largely on algorithmic design and improving recommender systems' performance (see Adomavicius and Tuzhilin 2005 for a review). Relatively little research has explored the impact of recommender systems on consumer behavior and decision making. Considering how important recommender systems have become as decision aids in online commerce, there is a need to explore the influence of these systems on consumer behavior.

We investigate the following research question: To what extent, if at all, do personalized recommender system ratings (indicating a system's estimates of users' preferences for items) influence users' willingness to pay for items? In an ideal scenario, recommender systems should provide decision-relevant information to consumers without manipulating or contaminating their behavioral and economic preferences. However,

based on behavioral theories in economics and psychology, we hypothesize that online recommendations significantly pull a consumer's willingness to pay in the direction of the recommendation. We test this idea with three controlled behavioral experiments in the context of digital music sales.

In the first study, we investigate whether randomly generated recommendations (i.e., not based on users' preferences) significantly impact consumers' willingness to pay for digital songs. In the second study, we extend these results by studying recommendation system error. We test whether the effects still exist for perturbations of actual recommendations generated by a real-time system that employs a popular, widely used recommendation algorithm. The third study investigates judgments made following mandatory consumption of song samples prior to indicating willingness to pay for songs. The requirement of mandatory consumption reduces memory-based uncertainty effects, allowing us to focus on processes at the time of preference formation rather than on recall processes tied to the earlier formation of preferences.

From a theoretical standpoint, the results of these studies contribute to our understanding of consumer behavioral economics and a heretofore unstudied influence of personalized recommendations on economic behavior, as well as inform the mechanisms by which the effect occurs. Furthermore, personalized recommendation systems clearly are prevalent decision aids in the online marketspace, and the unintended influences of these recommendations on consumers' judgments have significant practical implications.

2. Literature Review and Theory Development

2.1. Broader Perspective for the Study: Effects of Item Quality Information

In modern online settings, users receive a lot of information to help them make better decisions about which products to purchase or consume. In addition to *item content* information, *item quality* information is increasingly provided to consumers. Item quality information can take a variety of forms in online settings, sometimes all on the same site. It is, therefore,

useful to place our current research on personalized recommendations within the broader scope of item quality information and its impact on various outcomes of interest. Table 1 frames this broader scope along two dimensions: the type of item quality information provided to consumers and the level of granularity at which the impact of the information is measured.

First, item quality information can either be personalized or nonpersonalized. Personalized item quality information is typically generated using recommender systems that employ collaborative filtering and/or other techniques (Ricci et al. 2015) to provide recommendations to consumers that are tailored to the individual's preferences. Personalized item quality information can take the form of a predicted preference rating (e.g., we predict that you will like this movie as 4.5 stars out of 5) or simply a list of items predicted to be relevant for a given individual (e.g., here are top-5 movies specifically for you). Thus, personalized item quality information is likely to be very different from consumer to consumer. On the other hand, nonpersonalized, general item quality information represents an aggregate opinion (i.e., some population-level consensus) that is the same for all users. Common examples include average peer ratings (e.g., the average user rating for this movie is 4.5 stars out of 5), or popularity scales and lists derived from aggregating sales, downloads, or click data (e.g., here are top-5 best-selling movies for the past week).

The second dimension on which we frame the current research is the level of granularity at which the impact of item quality information is measured and analyzed (Table 1). At one end, one can adopt a micro-level perspective of how item quality information impacts the behaviors of individual consumers. Alternatively, one can adopt a macrolevel perspective that operates at a more aggregate level, e.g., the impact on total sales, downloads, or other aggregate outcomes of interest to retailers and the market in general.

The study of market-level outcomes has been an active research area, particularly on the effects of providing aggregate, nonpersonalized item quality information (Cell 4 of Table 1). Examples include Tucker and Zhang (2011), who studied the impact of popular

Table 1. Framework for Categorizing Research on Item Quality Information and Its Impact

Item quality information	Level of granularity at which impact is analyzed	
	Individual (micro-level)	Market (macro-level)
Personalized	1. Individual effects of a personalized system; e.g., how does a personalized recommendation influence the preference rating that a consumer states?	3. Market effects of a personalized system; e.g., how does Netflix's supplying of personalized ratings impact their revenues?
Nonpersonalized/ Aggregate	2. Individual effects of aggregate ratings; e.g., does knowing what others think affect a consumer's own judgment of an item?	4. Market effects of aggregate ratings; e.g., does showing consumers what others are buying lead to a convergence of sales for the highlighted items?

bestseller listings based on previous clicks upon the number of future clicks received. Along similar lines, Zhang and Liu (2012) studied how providing lenders with aggregate social information, i.e., in terms of the amount of funding already received, impacts future funding at an aggregate, market level. In the domain of digital goods, Godinho de Matos et al. (2016) investigated the impact of peer ratings, expressed as a list of most popular movies, on sales within a natural field experiment. Salganik et al. (2006) created a music market using a database of unknown songs and artists. They used an experimental methodology to manipulate whether or not the participants saw the number of downloads made by others, and then they measured the effect of this social influence on market factors. At the consumer level (Cell 2 of Table 1), Salganik and Watts (2008) used a similar design to support the effect that aggregated popularity feedback had on individual-level responses.

Although they deal with item quality information and its effects, the studies represented by Cells 2 and 4 are quite distinct from our own interest. These studies investigate more aggregate, nonpersonalized, and socially grounded types of information. The underlying theoretical mechanisms in Cells 2 and 4 have a clear social component; the hypotheses and implications are grounded in the literature on social influence. At an individual level, the general dynamic is one in which the consumer engages in a form of observational learning of how to behave based on observing the behavior of others. At the market level, in that the user is one of the "others" for other consumers, the behavior can create a self-reinforcing dynamic leading to market effects.

Personalized item quality information, such as personalized recommendations, by contrast, do not have an obvious social component. Depending on the algorithm, the personalized preference rating may or may not have any connection with others' behavior. Content-based algorithms, for example, depend on matching feature characteristics, not on the preferences of other users (Ricci et al. 2015). Even for algorithms that do incorporate others' preferences, e.g., collaborative filtering techniques (Ricci et al. 2015), the recommendation is presented as a system-generated rating that typically includes no explicit information about other users' preferences. The preference information that is submitted by users to a recommender system (as feedback to the system following item consumption) is typically private as well, i.e., not visible to other users. Thus, rather than mechanisms grounded in social psychology, the effects of personalized recommendations are posited on other bases (as discussed in Section 2.2).

Within the study of personalized approaches to recommendations, the research and development arms of retailers have rightly been interested in the market implications (Cell 3 of Table 1). An example of research in

this area views the role that personalized recommender systems play in promoting or inhibiting the diversity in individuals' choices. Fleder and Hosanagar (2009) used theoretical modeling and simulation to investigate how recommender systems impact consumer choices in aggregate. Their models suggest that recommender systems can lead to a rich-get-richer effect for popular products, resulting in a decrease in market-level sales diversity. In Hosanagar et al. (2014), the authors study whether recommender systems and online personalization lead to a fragmentation of the online population. Their results suggest that, surprisingly, personalization technologies such as recommender systems help users widen their interests, increasing the likelihood of commonality with others. Although these studies share an interest in personalized recommendations with the present study, the effects studied are generally at the market level and are not specifically concerned with understanding the biasing effects on individual consumers' economic decision making. The focus of our work is on the consumer side impacts of personalized recommendations (Cell 1 of Table 1). In this, the work is unique, providing a valuable contribution to furthering our understanding of an important and less-studied set of phenomena.

2.2. Personalized Recommender Systems' Effects on Preference Ratings

Personalized recommender systems can be viewed as operating within a feedback loop, as illustrated in Figure 1. Most commercial recommender systems (e.g., Netflix, iTunes, Amazon) take consumers' reported preferences as inputs for building predictive models (using methodologies from statistics, data mining, or machine learning) to systematically estimate consumers' ratings for not-yet-experienced items. In the most common setting, consumers represent their preferences in the form of self-reported, numeric ratings for items they have experienced. The system then delivers estimated ratings as personalized recommendations to help consumers in item-selection decisions. Consumers' subsequently submit ratings on newly consumed items as additional input for the system to adjust predictive models and estimate future recommendations. The consumer-reported preference ratings are also used to evaluate the recommender system's accuracy by comparing how closely the system-predicted ratings match the users' reported ratings.

There has been a significant amount of research on the design and implementation of recommender system algorithms, with the goal of improving accuracy performance (see Adomavicius and Tuzhilin 2005). However, only a handful of studies have considered the impact of personalized recommender-system-predicted ratings on consumer behavior and decision making (Cell 1 of Table 1).

Figure 1. Ratings as Part of a Feedback Loop in Consumer-Recommender Interactions
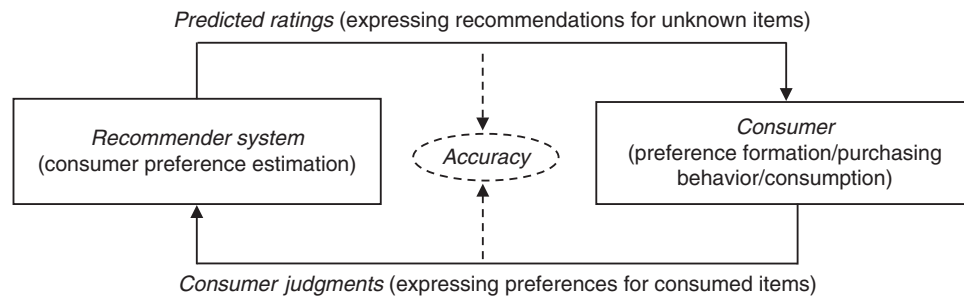
Source. Adapted from Adomavicius et al. (2013).

Cosley et al. (2003) explored the effects of personalized recommendations on user re-ratings of movies that they had seen in the past. The re-ratings tended to be highly consistent with the original ratings when users were asked to re-rate a movie with no preference prediction (i.e., recommendation) provided. However, when users were asked to re-rate a movie while being shown a predicted rating that was altered upward or downward from their original rating by one rating point, they tended to give higher or lower ratings, respectively. By contrast, Adomavicius et al. (2013) used designs for which preference ratings were elicited at the time of consumption. Even without a delay between consumption and elicited preference (i.e., without any recall-related uncertainty), systematic and substantial effects of system-generated recommendations were observed. The predicted rating, when perturbed to be higher or lower, affected the subsequently submitted consumer rating to move in the same direction.

Cosley et al. (2003) and Adomavicius et al. (2013) are the only studies to our knowledge that explore the impact of personalized recommender systems on consumers' preference judgments. The papers provide significant insights on how recommender systems impact preference construction; however, there are also significant limitations. First, it is not directly apparent how the effects of personalized recommender systems on consumers' self-reported preferences translate into consumers' real economic decision making: Do these effects transfer into consumer willingness-to-pay for a product and, ultimately, purchasing decisions? Even though personalized recommendations can impact a consumer's preference ratings as reported to a retail platform, these impacted responses may have little tangible effect on the consumer. However, effects of personalized recommendations on willingness-to-pay and purchasing decisions would represent real tangible consequences to the consumer and retailer in the form of surplus and economic welfare. Second, the participants in the Cosley et al. (2003) and Adomavicius et al. (2013) experiments had no direct incentive for accurately reporting their preferences. In the real world,

there is an underlying incentive for consumers to be as accurate as possible when reporting preferences because of their ongoing use and interaction with a personalized recommender system that incorporates those preferences. Reporting inaccurate preferences in real-world settings may reduce the value derived from the recommender system in future use. From the current experimental evidence, the question still remains as to whether the effect of system recommendations on preferences persists for properly incentivized participants. We address these limitations by measuring the effects of personalized recommendations on participants' willingness-to-pay in incentive-compatible, binding purchasing decisions.

2.3. Mechanisms for the Effects of Personalized Recommender Systems

As noted in Section 2.1, the research on general, non-personalized information on item quality is grounded in social influence mechanisms, in which the consumer is hypothesized to engage in a form of observational learning based on others' behaviors. Personalized recommendations typically do not share this social component and, thus, must derive from other mechanisms.

One theoretical lens for the research on the effects of system recommendations on preference judgments is anchoring and adjustment (Tversky and Kahneman 1974), an effect whereby a response is observed to be tilted toward an initial value (anchor). For example, when Tversky and Kahneman asked participants to guess the percentage of African countries that were members of the United Nations, those who were shown a random initial value of 10 gave a median estimate that 25% of countries in the UN are African; but if the random initial value was 65, the participants gave a median estimate that 45% of countries in the UN are African. Most of this research has, like this example, involved participants responding to questions of objective fact (see the review by Chapman and Johnson 2002). Only a few papers in the mainstream decision theory literature have looked directly at anchoring effects in preference construction, a situation where no

objective standard is available. Two examples are studied is reported by Schkade and Johnson (1989) and Ariely et al. (2003); however, their work studied preferences in abstract settings. There has also been little behavioral economics research on anchoring effects with consumer pricing behavior (i.e., willingness to pay). One example is the work by Ariely et al. (2006), who found that anchors based on social security numbers impacted participants' bids for items in an auction. Research has also found that charities can influence the amount that donors give by manipulating how donation options are presented; people will give more if the suggested options are higher (Surowiecki 2004, p. xxi). Our research moves into these less-charted areas.

As noted by numerous researchers, anchoring-and-adjustment describes an effect; it does not describe a particular mechanism by which the effect arises (see the review by Chapman and Johnson 2002). More generally, anchoring can be characterized as a specific example of judgments constructed in real time while being influenced by elements of the environment (Lichtenstein and Slovic 2006). Kahneman (2011) describes this as the WYSIATI (what you see is all there is) feature of automatic, involuntary thinking. From this view, as noted by Ariely et al. (2003), we might expect personalized system-displayed ratings to impact preferences in a way that is similar to the effects seen with questions of objective fact, though the general theory is still somewhat vague as to providing an explanation.

The general theory does allow us to posit several possible mechanisms whereby system recommendations might impact preference judgments. These are not intended as competing explanations. Indeed, anchoring (or focalism) effects have been consistently observed across many settings (see Epley and Gilovich 2010 for a review). The robustness of the effects may well be due to different explanations applying in different settings. Identifying these explanations grounds our work into a theory that benefits research in a couple of ways. First, they provide bases for why personalized system recommendations may or may not impact willingness-to-pay judgments. Second, they allow us to introduce manipulations and measures that inform us about the operation of different mechanisms whereby system recommendations may or may not have such influence. Based on existing findings, we identify three potential mechanisms by which system recommendations might impact consumers' stated preferences.

One explanation is a form of compatibility effect. In this mechanism, the numerical format of the information is the basis of the effect. Use of the same numerical scale for both the recommendation and the consumer's rating promotes a matching of the two, leading to an insufficient adjustment from the system-generated value. Specifically, the scale compatibility hypothesis (Tversky et al. 1988) argues that the more compatible the scales are (e.g., both measured in star-rating points on a 1–5 scale), the higher the weight of the recommendation in the decision process. The Tversky and Kahneman (1974) original hypothesis in terms of an anchoring-and-adjustment heuristic is this form of explanation (also see Wilson et al. 1996).

Another explanation is that system recommendation effects arise from uncertainty about one's preferences (e.g., Jacowitz and Kahneman 1995). This uncertainty can be characterized as a distribution or range of uncertain values. For example, if the preference construction is far removed in time from the experience being judged, the consumer may have some uncertainty about how much they liked the item from that distant experience of it. In forming a judgment via this mechanism, the person searches from the initial impression provided by the system recommendation to a nearby plausible value in the distribution or range of uncertain values, leading to final estimates tilted toward the recommendation.

Finally, there is an information-integration explanation. In situations involving objective facts, an anchor can be perceived as a suggestion provided by the context as to the correct answer (e.g., Northcraft and Neale 1987, Chapman and Bornstein 1996, Epley and Gilovich 2010). In our preference setting, this mechanism posits that the system recommendation can similarly be perceived as a piece of information to be used in the process of constructing a preference judgment. The mechanism can lead to use of the recommendation as information even where it is arguably not relevant (cf. Mussweiler and Strack 1999). Adomavicius et al. (2013) found support for this mechanism with effects on users' preference ratings.

2.4. Beyond Recommender Systems Effects on Preference Ratings: Hypotheses

Our work extends prior research in several important ways. In addressing the impacts of recommender systems on preferences and consumer behavior, we go beyond the extensive literature regarding the related anchoring effects. These past studies have largely been performed in artificial experimental settings not grounded in real-world practices and using tasks for which a verifiable outcome is being judged, leading to a bias measured against an objective performance standard (e.g., Chapman and Johnson 2002). By its very nature and from a practical standpoint, the recommendation setting of our research goes beyond this past research along both of these dimensions. Our research involves behavior in a realistic setting with subjective preferences and economic behavior that has clear monetary consequences to the individuals. Our participants are providing willingness-to-pay judgments for songs, and these judgments can result in actual song purchases.

More specifically, we extend the results of prior research on personalized recommendations as well. Beyond the effects of online recommendations on consumers' preference ratings, we hypothesize that personalized recommendations also impact consumers' willingness to pay for those items. Furthermore, we extend the research by addressing the possible mechanisms behind such effects. In our studies, the system ratings are provided on a 1–5 point scale as is commonly employed in real settings; however, for willingness to pay for the songs in our studies, the response is on a currency scale ranging from 0–99 cents. Thus, to the extent that a numerical adjustment process due to scale compatibility is at work, the effect should not differentiate between low and high ratings, since both are low relative to the currency scale. By contrast, if personalized recommendations impact willingness-to-pay judgments, this would indicate that other mechanisms are at play.

Mechanisms based on preference uncertainty could also pertain, since there often is uncertainty at the time of purchase. When recalling preferences formed in the past, uncertainty could reasonably be a component of the process underlying an effect on prices. However, a characteristic of digital goods is that they afford the possibility of free sampling prior to purchase (e.g., Shapiro and Varian 1999). Consequently, many online retail sites offer this capability, such as the "Look Inside" feature on Amazon.com or 30-second to one-minute samples of music on iTunes. These samples are intended to sharply reduce the uncertainty of preferences at the point of sale. To the extent that uncertainty is instrumental in creating recommendation effects, sampling should remove the effect or at least drastically curtail it. Finally, information integration mechanisms could be at work as well. Recommender systems are becoming common elements of many online experiences. It is plausible that recommendations could be increasingly viewed by consumers as highly relevant information and, thus, impact consumer preferences and economic behavior.

Aside from these mechanisms, another factor involved in using currency in our studies (and having participants make judgments with real monetary consequences) is the possible role of incentives. The evidence on the impact of incentives on behavioral phenomena is somewhat mixed. As reviewed by Camerer and Hogarth (1999), there is no evidence that rationality violations disappear with incentives; however, performance on effort-sensitive tasks can be affected. Anchoring effects have been shown to be sensitive to incentives for performance in a task with objective answers (Wright and Anderson 1989; cf. Müller et al. 2012); however, this may or may not extend to a preference setting where performance is not objectively measurable.

With these motivations in mind, the three studies address the following specific research hypotheses. The overall goal of the studies is to determine whether, and to what extent, the effects created by online personalized recommendations on preference ratings extend to impact consumers' pricing behavior. Additionally, we investigate the underlying mechanisms that are most likely responsible for any observed effects. To the extent that preference uncertainty and/or information-integration processes are at play, we expect consumers receiving an artificially or systematically inaccurate, low, or high recommendation to express a willingness to pay more in the direction of the corresponding recommendation. If the effect of system recommendations is largely a numerical adjustment process arising from scale compatibility effects, this would not be the case. The central hypothesis of Study 1 is stated in line with the former expectation as follows:

Hypothesis 1 (H1). *Participants exposed to randomly generated artificially high (low) recommendations for a product will exhibit a higher (lower) willingness to pay for that product.*

A common issue for recommender systems is error (often measured by root mean square error, RMSE) in predicting user ratings. Reducing prediction error has always been a key issue in recommender systems, for example, Netflix held a competition for a better recommendation algorithm with the goal of reducing prediction error, measured by RMSE, by 10% (Bennet and Lanning 2007). Recent studies demonstrated that inaccuracies (inadvertent or purposeful) in recommendations result in significant biases in consumers' preference ratings (Cosley et al. 2003, Adomavicius et al. 2013). We explore the economic impact of prediction error by introducing systematic perturbations into real system recommendations (based on a state-of-the-art recommender system algorithm). In this way, Study 2 controls for preference differences up front as part of the design, unlike in Study 1 where the recommendations are randomly generated. The procedure also allows us to observe how the effect varies as the magnitude of the perturbations is increased. The twofold design of Studies 1 and 2 parallels the setups employed by Adomavicius et al. (2013). In summary, the design of Study 2 tests for similar effects as in Study 1, but in a more realistic setting and with a more direct control of individual differences in user preferences (i.e., by using actual system-generated, real-time recommendations) and of the system's prediction error (i.e., by explicitly manipulating the size of systematic perturbations). The hypothesis is stated in a form parallel to that of Hypothesis 1:

Hypothesis 2 (H2). *Participants exposed to a recommendation that contains systematically introduced error in an*

upward (downward) direction will exhibit a higher (lower) willingness to pay for the product. The effect will increase with the magnitude of the introduced error.

Finally, Study 3 investigates the role of uncertainty as an explanation for recommendation effects. To the extent that uncertainty is instrumental in creating recommendation effects, sampling should remove the effect, or at least drastically curtail it. Study 3 addresses this explanation by forcing participants to sample each song prior to providing the willingness-to-pay preference judgment. The relevant hypothesis is stated in the form of the alternative hypothesis to be tested:

Hypothesis 3 (H3). *Participants forced to sample songs prior to stating their willingness to pay, when exposed to randomly generated artificially high (low) recommendations for a product, will exhibit a higher (lower) willingness to pay for that product.*

Studying how the impact of personalized system recommendations extends to willingness-to-pay judgments makes several contributions. Foremost, it furthers our understanding of the impact of personalized recommendations on willingness-to-pay judgments. Willingness-to-pay judgments have several differences from preference ratings (and from estimates of factual quantities) that relate to the possible occurrence of, and explanations for, the effects precipitated by personalized recommendations. Willingness-to-pay judgments are along a different scale than the system recommendations that are provided on a 1–5 point scale. Studying whether the effect is obtained under these conditions tests scale compatibility as an explanation. We use sampling tools, commonly available in consumer settings, as a means of checking explanations based on uncertainty. We have users experience artificial and variously perturbed recommendations as a means of checking information-integration explanations. Furthermore, the studies involve real economic behavior. Our participants make decisions about songs that will result in real purchases using personalized recommender systems that are already an important component of online sales.

3. Study 1: Recommendations and Willingness to Pay

Study 1 was designed to test Hypothesis 1 and establish whether randomly generated recommendations could significantly impact a consumer's willingness to pay (WTP). Participants provided judgments for popular songs that were offered for purchase in digital format during the study. Music was judged to be an appropriate stimulus as it is an experience good, for which users tend to have individualized, taste-driven preferences. The subject population used in all three of our studies were college-age individuals, because teens and young adults are a prime target audience for popular music. Music is also a good stimulus for our pricing experiments. Even with the free music streaming services and digital piracy, there is still a large market for purchased music. For example, PricewaterhouseCoopers projects a growth in the total U.S. music revenue to $18.04 billion by 2020.[3] Therefore, music provides a good context for our experiments.

3.1. Methods

3.1.1. Participants. We conducted a within-subjects study in the spring of 2012 at a university's behavioral research lab, with participants recruited from a college's research participant pool. Participants were paid a $10 fee plus a $5 endowment from which to purchase songs, as was described to them in the experimental procedure (discussed in Section 3.1.4). Seven participants were dropped from Study 1 because of response issues: three subjects had zero variance in their WTP judgments, and four subjects were outliers in terms of age relative to the desired subject population.[4] The final sample set consisted of 42 participants for analysis.

Demographic features of the sample are summarized in the first data column of Table 2. The participants were generally knowledgeable about music in the sense of our study, i.e., about evaluating and purchasing songs as a consumer good. Two-thirds (28/42) of the participants indicated buying music at least once a month, with only seven stating that they never buy music. Nearly three-quarters (31/42) of the participants said they owned more than 100 songs, with half (21/42) saying they own more than 1,000 songs, and only one participant indicated that they own no songs.

3.1.2. Stimuli. The stimuli were drawn from a database of 200 popular songs. The database consisted of songs in the bottom half of the year-end *Billboard* Hot 100 charts from 2006, 2007, 2008, and 2009 (i.e., 50 songs from each of these four years).[5] These charts provide a mix of popular songs that we expected would be viewed favorably by the participants and that they would be willing to purchase. The bottom half of these charts were used because we were interested in WTP judgments only for songs that the participants did not already own (thereby removing current ownership as an influence on the judgments). It was expected that songs in the bottom half of the charts would be more likely to be both favorable[6] and nonowned.

3.1.3. Willingness-to-Pay Judgments. To capture consumers' willingness to pay, we employed the incentive-compatible Becker–Degroot–Marshak (BDM 1964) method commonly used in experimental economics. For each song presented during the third task of the study,[7] participants were asked to declare a price they were willing to pay between zero and 99 cents. Participants were informed that five songs selected at random

Table 2. Participant Summary Statistics

	Study 1	Study 2	Study 3
No. of participants (n)	42	55	72
Average age, years (SD)	21.5 (1.95)	22.9 (2.44)	21.9 (2.77)
Gender	28 F, 14 M	31 F, 24 M	39 F, 33 M
Prior experience with recommender systems	50% (21/42)	47.3% (26/55)	77.8% (56/72)
Student level	36 undergrad, 6 grad	27 undergrad, 25 grad, 3 other	49 undergrad, 22 grad, 1 other
How often do you buy music?			
Never	7	4	13
Once a year	7	15	27
Once a month	23	31	24
Once a week	3	3	6
More than twice a week	2	2	2
Number of songs owned			
None	1	1	0
1–100	10	9	12
101–1,000	10	19	35
1,001–10,000	19	23	19
More than 10,000	2	3	6

at the end of the study would be assigned random selling prices, based on a uniform distribution, between 1 and 99 cents. For each of these five songs, the participant was required to purchase the song using money from their $5 endowment at the randomly assigned selling price if it was equal to or below their declared willingness to pay. The participant would keep the remainder of the endowment. This procedure is akin to real pricing decisions: If the selling price is at or below the consumer's value, the selling price is paid; if it is not, the purchase is not made. Prior to making any willingness to pay judgments, we presented participants with a detailed explanation of the BDM method so that they understood how the procedure incents accurate reporting of their willingness to pay. Participants took a quiz to check their knowledge of the procedure: 35/42 (83%) of participants answered the quiz questions perfectly on the first pass. Feedback was provided to correct mistaken answers before proceeding with the pricing task within the following procedure.

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

Task 1. Each participant was asked to provide preference ratings for at least 50 songs randomly selected from the aforementioned pool of 200 songs. Ratings were provided using a scale from one to five stars with half-star increments, having the following verbal labels: *, "Hate it;" **, "Don't like it;" ***, "Like it;" ****, "Really like

it;" and *****, "Love it."[8] For each song, the artist name(s), song title, duration, album name, and a 30-second sample were provided. Participants could choose to listen to any song at any time by clicking the link of the song sample. The objective of the song-rating task was to capture music preferences of the participants. This task provided a seeming basis for the system recommendations provided later in Study 1 (although the ratings were not used for this purpose) and as a real basis for the system recommendations in Study 2. These ratings also provide a basis for the post hoc analysis of Study 1 to be discussed in Section 3.2.

Task 2. A different list of songs was presented (with the same information for each song as in the first task and with 20 songs on each screen), again randomly drawn for each participant from the same set of 200 songs but excluding the songs rated by the participant in Task 1. For each song, the participant was asked whether or not they owned the song until 40 nonowned songs were identified.[9] Songs that were owned were excluded from the third task, in which willingness-to-pay judgments were obtained.

Task 3. Participants first underwent training for the BDM pricing method as described in Section 3.1.3. They then completed a within-subjects experiment where the treatment was the star rating of the song recommendation and the dependent variable was the willingness to pay. Participants were presented with 40 songs that they did not own (identified during the second task), along with a star-rating recommendation, artist name(s), song title, duration, album name, and a 30-second sample for each song. The songs were randomly ordered and assigned to the different manipulation conditions. Participants were asked to specify

their willingness to pay for each song on a scale from 0 to 99 cents. The star rating recommendations were presented as personalized ratings for each participant. All recommendations were presented with a one decimal place precision numerical value (e.g., 4.3 stars) and with a graphical star representation of the same value. Ten of the 40 songs presented to each participant had a randomly generated *low* recommendation between 1.0 and 2.0 stars drawn from a uniform distribution (*Low* condition), 10 were presented with a randomly generated *high* recommendation between 4.0 and 5.0 stars (*High* condition), 10 were presented with a randomly generated *mid*-range recommendation between 2.5 and 3.5 stars (*Mid* condition), and 10 were presented with *no* recommendation to act as a control (*Control* condition). The 30 songs presented with recommendations were randomly ordered and presented together on one webpage. The 10 control songs were presented on the following webpage. The separation allowed us to include instructions that these were songs for which "our system did not have enough data to make predictions," thus providing an explanation for why no recommendations accompanied these items.

***Task* 4.** Participants completed a short survey providing demographic and other individual information. Next, we drew random prices for five randomly chosen songs from Task 3, and purchases were automatically enforced as described in Section 3.1.3. The participation fee and the endowment, minus prices paid for the required purchases, were distributed to participants in cash. MP3 versions of the songs purchased by participants were gifted through Amazon.com within 12 hours of the study's conclusion.

3.2. Results
The distribution of willingness-to-pay judgments overall is shown in Figure 2. As expected, a distribution with a right skew is observed. The overall mean was 22.7 cents (SD = 25.1) with 28.7% (471/1,640) of the song pricing observations having a stated value of 0,

Figure 2. Summary Distributions of Willingness-to-Pay Judgments for Studies 1–3
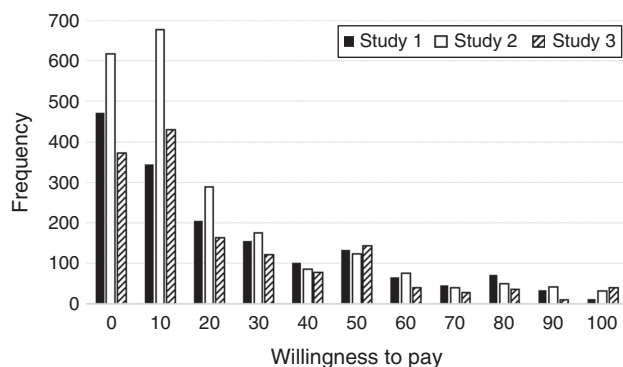
Figure 3. Average Willingness-to-Pay by Treatment Group in Study 1
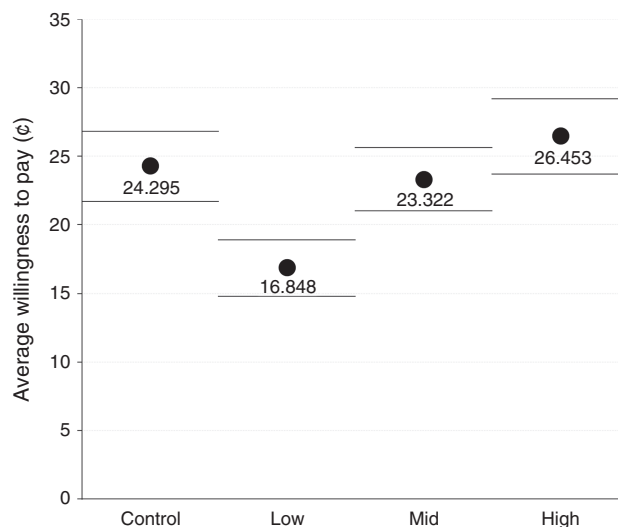
Note. Error bars represent 95% confidence intervals around mean of observations.

coming from 57.1% (24/42) of the subjects. So, subjects were quite willing to value songs at 0 when they saw no value to a song; the study did not provide an overwhelming demand for positive responses. These values support that subjects did realize the impact of positive responses, and avoided them when judged appropriate.

Figure 3 shows a plot of the aggregate means of willingness to pay for each treatment group. We performed a repeated measure analysis of variance (ANOVA) to test for differences across the three main treatment levels: High, Mid, and Low. With repeated measures ANOVA, adjustments are needed if the variance-covariance matrix of the dependent variables indicates significant differences in variances between conditions (Mauchly 1940). We utilized the most conservative adjustment factor—Box's conservative epsilon—to correct for violations of the sphericity assumption (Box 1954a, b). The adjusted analysis demonstrated a statistically significant effect of the shown rating on willingness to pay ($F(2, 1196) = 42.27$, $p < 0.001$); so, we followed with pairwise contrasts using t-tests, Monte Carlo Fisher–Pitman permutation tests, and the Wilcoxon signed-rank tests across treatment levels and against the control group as shown in Table 3. All three main treatment conditions significantly differed, showing a clear positive effect of the treatment on pricing behavior. The control group demonstrated an intermediate willingness to pay, showing a statistically significant difference from the Low treatment condition but not from the High treatment condition (one-tailed $p < 0.0001$ and $p = 0.13$, respectively) and no difference from the Medium condition (two-tailed $p = 0.58$).

Table 3. Pairwise Comparison Test Values and Significance Levels for Comparisons of Aggregate Treatment Group Means

	Control	Low	Mid
t-tests			
Low (1–2 stars)	5.3805***		
Mid (2.5–3.5 stars)	0.8175	5.1996***	
High (4–5 stars)	1.1373	4.5258***	2.1438*
Fisher–Pitman permutation test			
Low (1–2 stars)	344.3***		
Mid (2.5–3.5 stars)	58.1	292.3***	
High (4–5 stars)	87.9	444.7***	156.5*
Wilcoxon signed-rank test			
Low (1–2 stars)	4.288***		
Mid (2.5–3.5 stars)	0.511	4.302***	
High (4–5 stars)	0.968	4.232***	1.400

Note. Two-tailed tests for Control versus Mid, all other contrasts were directional and tested with 1-tailed tests.

$^*p \leq 0.05$; $^{***}p \leq 0.001$.

Next we performed a series of post hoc regression analyses to further explore the relationship between the shown star rating and willingness to pay, while controlling for participant-level factors. Although there were three treatment groups, the actual ratings shown to participants were randomly assigned star ratings from within the range for the corresponding treatment group (*Low*: 1.0–2.0 stars, *Mid*: 2.5–3.5 stars, *High*: 4.0–5.0 stars). Thus, in the analysis, the shown rating was a continuous variable ranging from 1.0–5.0 and was the main independent variable of interest. Control variables for demographic and consumer-related factors were included, as collected by the survey in Task 4. Additionally, we controlled for the participants' preferences by calculating a predicted star rating recommendation for each song (on a 5-star scale with one decimal precision), using the popular and widely-used item-based collaborative filtering algorithm (IBCF) (Sarwar et al. 2001). Several recommendation algorithms were evaluated based on the data from Task 1 of Study 1, and the IBCF technique was selected in all three studies because it had the highest predictive accuracy. More details on the IBCF recommendation algorithm and its performance are presented in Online Appendix 2. In Study 2, the system recommendations are an instrumental part of the manipulations in the study design. Here, in Study 1, we use IBCF to control for the system's estimation of consumer preferences as a proxy of consumer's preferences. By including this predicted rating (which was not shown to any participant during Study 1) in our analysis, we are able to determine if the random system-displayed ratings had an impact on willingness to pay above and beyond the participant's predicted preferences.

The resulting baseline regression model is shown below. In the model, WTP_{ij} is the reported willingness to pay for participant i on song j, $ShownRating_{ij}$ is the recommendation star-rating shown to participant i for song j, and $PredictedRating_{ij}$ is the predicted recommendation star-rating for participant i on song j (i.e., the predicted preference). In addition, $Controls_i$ is a vector of demographic and consumer-related variables for participant i. The controls included in the model were as follows: gender (binary); age (integer); school level (undergrad: yes/no, binary); whether the participant has prior experience with recommendation systems (yes/no, binary); preference ratings (interval five-point scale) for the music genres country, rock, hip hop, and pop; the number of songs owned (ordinal five-point scale, representing "none," "1–100," "101–1,000," "1,001–10,000," and "more than 10,000"); frequency of music purchases (ordinal five-point scale, representing "never," "once a year," "once a month," "once a week," and "more than twice a week"); whether they thought recommendations in the study were accurate (interval five-point scale); and whether they thought the recommendations were useful (interval five-point scale). It was not expected that participants would have exact values for their song collections and purchase frequency, so we used ordinal five-point scales to identify qualitative distinctions in music ownership and purchases. Since the study utilized a repeated measures design with a balanced number of observations on each participant, the composite error term ($u_i + \varepsilon_{ij}$) includes an individual participant effect u_i in addition to the standard disturbance term ε_{ij}

$$WTP_{ij} = b_0 + b_1(ShownRating_{ij}) + b_2(PredictedRating_{ij})$$
$$+ \mathbf{b}_3(Controls_i) + u_i + \varepsilon_{ij}.$$

Three regression models were estimated and compared to account for the nature of the dependent variable. The baseline regression model (Model 1 in Table 4) used generalized least squares (GLS) with random effects estimation and robust standard errors, clustering errors by participant. To control for participant-level heterogeneity, a random effect was used to model the individual participant effect. Random effects were chosen over fixed effects for three key reasons.[10] First, we assume that the effects of the participants are randomly drawn from the overall population of potential participants. Second, the results of a Hausman test deemed the random effects model appropriate. Third, using random effects allows us to include participant-level controls in the analysis.

Prices and willingness-to-pay judgments often follow a log-normal distribution (see Figure 2). Therefore, we estimated a log-normal GLS regression with random participant-level effects in Model 2, using $\ln(WTP + 1)$ as the dependent variable to account for the skewed distribution of willingness to pay. For Model 3, a Tobit regression with participant-level random effects was

Table 4. Study 1 Regression Results; Dependent Variable: Willingness to Pay

	Model 1: GLS, RE	Model 2: LogNorm, RE	Model 3: Tobit, RE
ShownRating	3.533 (0.80)***	0.162 (0.03)***	3.530 (0.39)***
PredictedRating	6.235 (1.72)***	0.396 (0.12)***	6.215 (1.11)***
Controls			
Male	−9.466 (4.29)*	−0.910 (0.29)*	−9.463 (5.26)
Undergrad	−3.819 (12.60)	−0.335 (0.71)	−3.826 (11.54)
Age	−1.366 (1.75)	−0.069 (0.12)	−1.362 (2.05)
usedRecSys	−12.437 (5.57)*	−1.152 (0.37)*	−12.439 (5.82)*
Country	2.608 (1.56)	0.204 (0.10)*	2.611 (1.91)
Rock	1.634 (2.74)	0.275 (0.17)	1.635 (2.94)
Hiphop	−0.468 (2.39)	0.076 (0.14)	−0.459 (2.39)
Pop	3.176 (2.42)	0.091 (0.18)	3.176 (2.95)
recomAccurate	−3.963 (3.67)	−0.238 (0.28)	−3.961 (4.97)
recomUseful	4.193 (3.53)	0.294 (0.27)	4.200 (4.27)
buyingFreq	1.921 (2.44)	0.227 (0.15)	1.927 (2.86)
songsOwned	−4.120 (3.83)	−0.276 (0.26)	−4.109 (4.04)
Constant	18.543 (52.21)	1.220 (3.40)	18.425 (56.98)
R^2	0.214	0.268	
χ^2	89.35***	142.18***	136.15***

Notes. Number of clusters = 42, $n = 1{,}240$ (42 participants × 30 responses—20 missing per endnote 9, for each analysis). Standard errors are in parentheses. All models use robust standard error estimation, clustered by participant. Model summaries: Model 1—GLS estimation with random participant-level effects (RE); Model 2—log-normal GLS (i.e., dependent variable = $\ln(WTP + 1)$) with random participant-level effects; Model 3—Tobit regression (upper limit 99, lower limit 0) with random participant-level effects. All models estimated using the Stata 14 software.

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

estimated to account for potential censoring of the dependent variable (i.e., the participants' responses for willingness to pay were limited to a maximum value of 99 and a minimum value of 0). Tobit models are commonly used to model willingness-to-pay (e.g., Donaldson et al. 1998); and, the assumption that zero-measured values of WTP reflect actual zero values and not a decision to enter the market holds in our controlled experimental setting.

Table 4 presents the estimated coefficients and standard errors for the three models, and the results are highly consistent for the primary independent variables of interest, suggesting robustness of the results. All models utilized robust standard error estimates and clustered errors by participant. Note that the control treatment observations (i.e., where no recommendation was provided) were not included in the regression analyses since they had null values for the main independent variable *ShownRating*.

The results of our analysis for Study 1 are strongly consistent with H1 and in line with the ANOVA. They clearly demonstrate a significant effect of shown recommendations on consumers' pricing behavior. We observed that randomly generated recommendations with no dependence on actual user preferences can impact consumers' willingness to pay for an item, while

controlling for participant-level factors and the participant's predicted preferences for the product being recommended.

From the log-normal model (Model 2), we observe an increase (decrease) of 17.6% (i.e., $\exp(0.162)$) in willingness to pay for each 1-star increase (decrease) in the shown recommendation rating. The GLS model provides similar results: a 1-star increase (decrease) in the shown recommendation results in a 3.53 cents U.S. increase (decrease) in willingness to pay, in a sample with a mean willingness to pay ≈ 22.7 cents U.S. The Tobit model provides similar results, although it should be noted that the coefficient for *ShownRating* (3.530, $p < 0.001$) represents the marginal effect on the unobserved latent variable y^*, the uncensored willingness to pay. Using the *margins* command in Stata, we computed the marginal effect for the conditional mean specification $E(\text{WTP} \mid \mathbf{x}, 0 \leq \text{WTP} \leq 99)$, where $\mathbf{x}$ represents the collection of independent variables. This adjusted marginal effect takes into account censoring and was observed to be a 2.86 cents U.S. ($p < 0.001$) increase (decrease) in willingness to pay for each 1-star increase (decrease) in the shown rating. Together, the regression results suggest that we can conservatively expect an increase of approximately 12%–17% in willingness to pay for each 1-star increase in shown rating.

4. Study 2: Errors in Personalized Recommendations

The goal of Study 2 was to extend the results of Study 1 by testing Hypothesis 2 on the effect of system errors. In this study, we explore the impact of systematically introduced errors of different magnitudes in recommendations on consumers' willingness to pay. The recommendations presented to participants were calculated during the experiment, thereby connecting the regression model to the experimental design using truly controlled independent variables. This way, the amount of introduced error is directly manipulated in the study for investigation of its effect in a controlled manner. The design of the study is intended to test for similar effects as in Study 1, but in a more realistic setting with system-generated, real-time recommendations and better control of a particular potential aspect of system's prediction error.

4.1. Methods
4.1.1. Participants. Participants in Study 2 used the same facilities and were recruited from the same pool and during the same time period as in Study 1 (i.e., spring 2012); however, there was no overlap in participants across the two studies. The same participation fee and endowment used in Study 1 was provided to participants in Study 2. Fourteen participants were dropped from Study 2 because of response issues: five subjects had zero variance in their WTP judgments

and nine subjects were outliers in terms of age relative to the desired subject population (see endnote 4). The final sample set consisted of 55 participants for analysis.

Demographic features of the sample are summarized in the second data column of Table 2. The participants are similar to those in Study 1 except for a lower percentage of undergraduates in the sample. Those in Study 2 were equally knowledgeable about purchasing music. Almost two-thirds (36/55) of the participants indicated buying music at least once a month, with only four stating that they never buy music. More than 80% (45/55) of the participants said they owned more than 100 songs, with nearly half (26/55) saying they own more than 1,000 songs; only one participant indicated that they own no songs.

4.1.2. Procedure. The stimuli database of 200 songs and the four tasks of the study were identical to Study 1. All participants completed the initial song-rating and song ownership tasks as in Study 1. Willingness-to-pay judgments were obtained using the BDM method with participant training, testing, and feedback (36/55 = 65% of participants answered the quiz questions perfectly on the first pass). The final survey, payouts, and song distribution were also conducted in the same manner as in Study 1. The mean duration for participants to complete the session was just under 27 minutes, similar in length to Study 1. The only difference between Studies 1 and 2 was in the design used for Task 3.

Study 2 used a within-subjects design with willingness to pay as the dependent variable. Unlike Study 1, the ratings that participants provided in Task 1 of Study 2 were used as inputs to a recommender system to calculate participant's preferences for the songs used in Tasks 2 and 3 in real time. Details on the IBCF recommendation algorithm used, and its performance, are presented in Online Appendix 2. The predicted ratings were then systematically perturbed (i.e., excess error was introduced to each recommendation) to generate the shown recommendation ratings. Perturbations of −1.5 stars, −1 star, −0.5 stars, 0 stars, +0.5 stars, +1 star, and +1.5 stars were added to the actual recommendations, representing seven treatment levels. Thus, Study 2 applied a systematic perturbation technique that better controlled for individual preference differences, allowing an investigation of the effects of introduced error on willingness to pay.

Each participant provided WTP judgments for 40 songs, five from each of the seven treatment levels and five controls. The 30 songs with perturbed ratings were determined pseudo-randomly to assure that the manipulated ratings would fit into the five-point rating scale. First, 10 songs with predicted rating scores between 2.5 and 3.5 were selected randomly to receive perturbations of −1.5 and +1.5. From the remaining songs, 10 songs with predicted rating scores between

2.0 and 4.0 were selected randomly to receive perturbations of −1.0 and +1.0. Then, 10 songs with predicted rating scores between 1.5 and 4.5 were selected randomly to receive perturbations of −0.5 and +0.5. Finally, five songs were randomly selected and their predicted ratings were not perturbed; they were displayed exactly as predicted. These 35 songs were randomly intermixed. Following this, a final set of five songs were added as a control in random order (i.e., with no predicted system rating provided and with the instructions that these were songs for which "our system did not have enough data to make predictions" as an explanation).
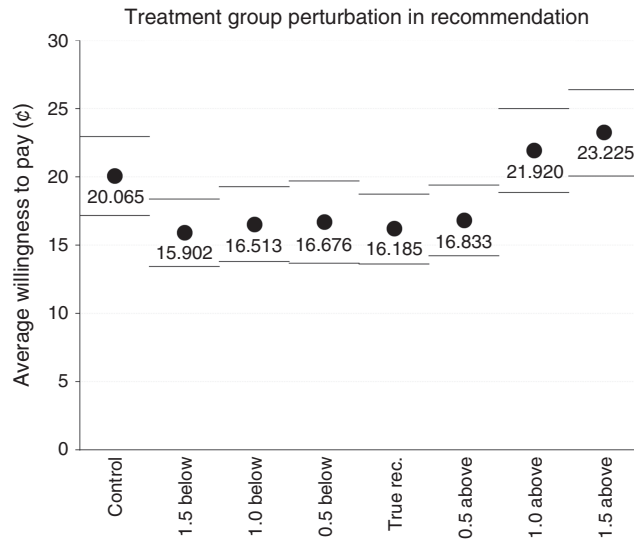
4.2. Results

Figure 2 shows the distribution of willingness-to-pay judgments overall, alongside those of the other studies. A similar right-skewed distribution is observed. The overall mean was 18.4 cents (SD = 23.9) with 28.0% (617/2,200) of the song pricing observations having a stated value of 0, coming from 58.2% (32/55) of the subjects. So, the responses were similar to those in Study 1, with subjects willing to value songs at 0 when they saw no value to a song, and with no observable experimentally induced demand for positive responses.

As in Study 1, we start with a repeated measures ANOVA. After correction for the violations of sphericity using Box's conservative epsilon (Box 1954a, b), the ANOVA confirmed a significant treatment effect of recommendation error (i.e., perturbation size) on willingness to pay (leaving out the control items, $F(6, 1864) = 8.17$, $p = 0.0045$). Figure 4 presents the aggregate means by treatment condition. Using planned contrasts, comparing the High to the Low perturbations at each level—1.5, 1, and 0.5—there is a significant difference in the responses at 1.5 and 1.0 between the High and Low groups ($p < 0.001$) but not at 0.5 ($p > 0.90$). Thus, there is no indication of a diminution in the effect as the size of the perturbation increases.

Comparing the control group to each of the other groups, the mean responses to the songs in the control condition (no recommendation was given) were significantly different at an $\alpha = 0.05$ level (two-tailed $t(274)$ not assuming equal variances) compared to mean responses in the True condition and the 1.5 Below condition.[11] They were marginally significantly different, at an $\alpha = 0.10$ level compared to the mean responses in the 1.0 Below condition. Thus, the negative perturbations tended to pull down willingness to pay compared to the control. However, for a more sensitive test and a clearer overall picture, we can apply regression modeling to the non-control conditions that vary in manipulations from −1.5 to +1.5.

Moving to our main analysis of Study 2, we used regression to examine the relationship between recommendation error and willingness to pay. We follow a

Figure 4. Average Willingness-to-Pay by Treatment Group in Study 2

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

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

	Model 1: GLS, RE	Model 2: LogNorm, RE	Model 3: Tobit, RE
Perturbation	2.245 (0.880)*	0.121 (0.049)*	2.245 (0.392)***
PredictedRating	8.787 (1.414)***	0.567 (0.082)***	8.797 (0.832)***
Controls			
Male	−2.203 (4.682)	0.108 (0.3)	−2.203 (4.253)
Undergrad	−4.814 (4.437)	−0.173 (0.297)	−4.814 (4.54)
Age	−0.351 (0.892)	−0.007 (0.055)	−0.351 (0.934)
usedRecSys	3.944 (4.115)	0.116 (0.272)	3.944 (4.057)
Country	−2.646 (1.535)	−0.166 (0.112)	−2.647 (2.091)
Rock	−1.92 (1.601)	−0.245 (0.101)*	−1.92 (1.977)
Hiphop	0.146 (2.032)	−0.046 (0.115)	0.146 (1.834)
Pop	−0.921 (1.838)	−0.065 (0.114)	−0.921 (2.221)
recomAccurate	1.089 (2.104)	0.081 (0.162)	1.089 (2.451)
recomUseful	3.886 (1.908)*	0.218 (0.131)	3.886 (1.902)*
buyingFreq	2.712 (1.742)	0.334 (0.106)**	2.711 (2.407)
songsOwned	−3.576 (2.573)	−0.139 (0.184)	−3.576 (2.647)
Constant	7.613 (23.674)	1.065 (1.564)	7.587 (25.841)
R^2	0.1580	0.1850	
χ^2	94.32***	176.94***	165.99***

Notes. Number of clusters = 55, $n = 1{,}925$ (55 participants × 35 responses, for each analysis). Standard errors are in parentheses, all models use robust standard error estimation, clustered by participant. Model summaries: Model 1—GLS estimation with random participant-level effects; Model 2—log-normal GLS (i.e., dependent variable = $\ln(WTP + 1)$) with random participant-level effects; Model 3—Tobit regression (upper limit 99, lower limit 0) with random participant-level effects. All models estimated using the Stata 14 software.
*$p \leq 0.05$; **$p \leq 0.01$; ***$p \leq 0.001$.

similar approach as in Study 1 and analyze the relationship using three regression models. The distribution of willingness-to-pay data in Study 2 was similar to that of Study 1, so the same analysis strategy was adopted. The main difference in this analysis as compared to Study 1 is that the primary explanatory variable is $Perturbation_{ij} = ShownRating_{ij} - PredictedRating_{ij}$ which captures the systematic error we introduced through treatments. The baseline model utilized GLS regression with random participant-level effect (the Hausman test for fixed versus random effects suggested that a random effects model was appropriate). We control for the participants' preferences using the predicted rating for each song in the study (i.e., the recommendation rating prior to perturbation), which was calculated using the IBCF recommendation algorithm. In this analysis, these variables are generated not post hoc; instead, they are part of the study design up front. The same set of control variables used in Study 1 was included in our regression model for Study 2. The resulting regression model is

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

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

As can be seen in Model 2 of Table 5, we observed a significant increase of approximately 12.9% (i.e.,

exp(0.121)) in WTP for each 1-star positive increase in error of the shown recommendation, based on the log-normal regression model. The GLS model shows a 2.245 cents U.S. ($p \leq 0.05$) increase in willingness to pay for each 1-star positive increase in error of the shown recommendation. The Tobit model provided similar results: We observed a marginal effect of 2.245 cents U.S. ($p < 0.001$) of the perturbation on the latent variable $y*$, which is the unobserved and uncensored willingness to pay. The marginal effect for the conditional mean specification $E(WTP \mid \mathbf{x}, 0 \leq WTP \leq 99)$, where $\mathbf{x}$ represents the collection of independent variables, is a 1.74 cents U.S. ($p < 0.001$) increase in willingness to pay for each 1-star increase in error of the shown recommendation. The sample mean willingness to pay for Study 2 is 18.4 cents U.S.; the regressions suggest that a significant effect corresponding to approximately 9.5%–12.9% change in willingness to pay for each 1-star of systematic error in the recommendation can be expected.

The results of Study 2 provide strong support for H2 and extend the results of Study 1 in two important ways. First, Study 2 provides more realism to the analysis, since it utilizes real recommendations generated using an actual real-time system that applies a popular, commonly used recommendation algorithm. Second, rather than randomly assigning recommendations as

in Study 1, in Study 2 the recommendations presented to participants were calculated based on their preferences and then perturbed to introduce varying levels of systematic error. Study 2 demonstrates the potential impact of these system errors. The study also shows that the magnitude of error introduced into the recommendations does not lead to a reduction in the impact of recommendations on willingness to pay. The effect of perturbation is significant and linear, with no indicated curvature.

5. Study 3: Reducing Uncertainty

Study 3 was designed to explore whether uncertainty is a critical component of the effect of system recommendations (Hypothesis 3). This study largely followed the design of Study 1 and provided randomly generated song recommendations, except that listening to song samples was mandatory for the participants before pricing the songs, and each priced song was displayed sequentially on a separate page.

5.1. Methods

5.1.1. Participants. Participants in Study 3 used the same facilities and were recruited from the same participant pool during the spring of 2014; there was no overlap in participants across the three studies. The same participation fee and endowment used in Study 1 was provided to participants in Study 3. Twelve participants were dropped from Study 3 because of response issues: one subject experienced a technical issue, two subjects had zero variance in their WTP judgments, and nine subjects were outliers in terms of age relative to the desired subject population (see endnote 4). The final sample set consisted of 72 participants for analysis.

Demographic features of the sample are summarized in the third data column of Table 2. The participants are similar to those in Studies 1 and 2 except there is a higher percentage of experience with recommender systems and an intermediate number of undergraduates compared to the other two samples. The Study 3 participants were equally knowledgeable about purchasing music. Almost four-fifths (58/72) of the participants indicated buying music at least once a month, with only 13 stating that they never buy music. More than four-fifths (61/72) of the participants said they owned more than 100 songs, with more than a third (26/72) saying they own more than 1,000 songs, and no one indicated that they own no songs.

5.1.2. Procedure. The stimuli database of 200 songs and the basic tasks of the study were identical to Studies 1 and 2. All participants completed the initial song-rating and song-ownership tasks as in those studies, except only 20 nonowned songs were identified in Task 2, instead of 40. The number of songs was reduced

to keep the experimental duration at a comparable time to the other studies. The mean duration for participants to complete the session was 30 minutes, so we were successful. After training with the BDM method, 51/72 = 71% of participants answered the quiz questions perfectly on the first pass. The final survey, payouts, and song distribution were also conducted in the same manner as before. The only difference between the studies was in the design used for Task 3; in this case, Task 3 incorporated a forced listening of song samples prior to the pricing task with each song presented separately.

Study 3 repeated the within-subjects design with willingness to pay as the dependent variable. In Task 3, five songs were presented with a randomly generated *low* recommendation between 1.0 and 2.0 stars drawn from a uniform distribution, five were presented with a randomly generated *high* recommendation between 4.0 and 5.0 stars, five were presented with a randomly generated *mid*-range recommendation between 2.5 and 3.5 stars, and five were presented with *no* recommendation to act as a control. Each song was presented separately on its own page along with the system recommendation, where applicable (see the sample page in Online Appendix 1). The sample began playing as soon as the page loaded and continued for 30 seconds. The participants were not allowed to enter their willingness-to-pay judgment or to proceed to the next page until the 30 seconds had elapsed. After that, participants had the option to play the song sample again by clicking on the song title, or to respond and move to the next song. The 15 songs presented with recommendations were randomly ordered and shown first. The five control songs were presented next with the instructions that these were songs for which "our system did not have enough data to make predictions" as an explanation. We recorded the order and the time stamp of each response, since each response was on a separate page for this study.

5.2. Results

Figure 2 illustrates a similar right-skewed distribution of willingness-to-pay judgments as observed in Studies 1 and 2. The overall mean was 20.36 cents (SD = 25.3), with 25.8% (372/1,439) of the song pricing observations having a stated value of 0, coming from 58.3% (42/72) of the subjects. As before, subjects showed no observable experimentally induced demand for positive responses.

Figure 5 presents the aggregate means by condition. After correcting for violations of the sphericity assumption using Box's conservative epsilon (Box 1954a, b), the repeated measures ANOVA confirmed that there was a significant treatment effect of recommendation on willingness to pay across the three treatment conditions: High, Medium, and Low ($F(2, 1006) = 8.04$, $p = 0.0048$). Pairwise contrasts based on t-tests, Monte Carlo Fisher–Pitman permutation tests, and Wilcoxon signed-rank

Figure 5. Average Willingness to Pay by Treatment Group in Study 3
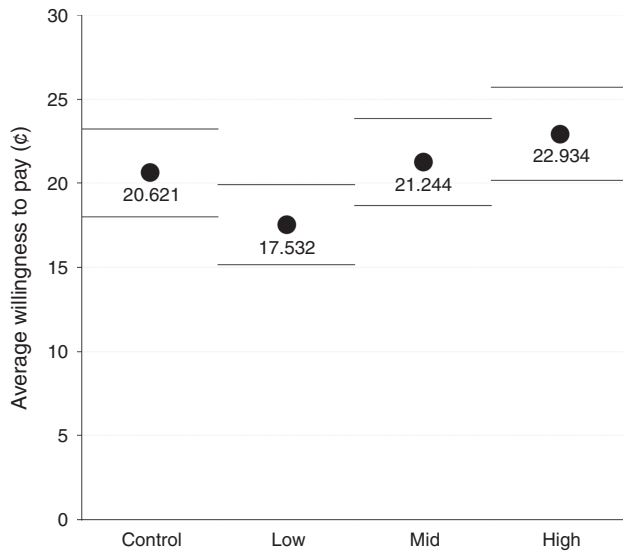
Note. Error bars represent 95% confidence intervals around the mean of observations.

tests (Table 6) indicated a difference between the Low group and the Medium, High, and Control groups.

Next we performed the same post hoc regression analysis applied in Study 1 with the following differences. Because we were able to record the order of the songs that each participant listened to in Study 3, we could include the display order of songs as an independent variable in the regression analysis (*DisplayOrder*, integer). This design aspect is useful as a check on whether any number of issues might have been operative over the course of the study session. We can test for order effects to identify if the impact of recommendations on willingness to pay diminishes

Table 6. Pairwise Comparison Test Values and Significance Levels for Comparisons of Aggregate Treatment Group Means

	Control	Low	Mid
t-tests			
Low (1–2 stars)	1.823*		
Mid (2.5–3.5 stars)	0.326	2.106*	
High (4–5 stars)	1.373	2.939**	1.340
Fisher–Pitman permutation test			
Low (1–2 stars)	210.05*		
Mid (2.5–3.5 stars)	35.75	245.8*	
High (4–5 stars)	158.95	369**	123.2
Wilcoxon signed-rank test			
Low (1–2 stars)	1.325		
Mid (2.5–3.5 stars)	0.485	2.026*	
High (4–5 stars)	0.915	2.708**	0.943

Note. Two-tailed *t*-test for Control versus Mid, all other contrasts were directional and tested with 1-tailed *t*-test.

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

Table 7. Study 3 Regression Results; Dependent Variable: Willingness to Pay

	Model 1: GLS, RE	Model 2: LogNorm, RE	Model 3: Tobit, RE
ShownRating	1.998 (0.65)*	0.1 (0.03)*	1.999 (0.48)***
PredictedRating	8.458 (1.43)***	0.604 (0.09)***	8.457 (1.12)***
DisplayOrder	0.119 (0.15)	−0.004 (0.01)	0.119 (0.12)
DisplayOrder× ShownRating	0.089 (0.1)	−0.002 (0.01)	0.09 (0.1)
Controls			
Male	−1.412 (4.53)	−0.129 (0.24)	−1.412 (4.15)
Undergrad	−0.837 (6.4)	0.278 (0.33)	−0.836 (5.95)
Age	−0.086 (0.97)	0.013 (0.06)	−0.086 (0.97)
usedRecSys	5.141 (4.57)	0.359 (0.26)	5.141 (4.77)
Country	3.935 (1.66)*	0.241 (0.1)*	3.935 (1.78)*
Rock	1.457 (2.45)	−0.036 (0.12)	1.457 (2.29)
Hiphop	0.18 (1.73)	−0.02 (0.11)	0.18 (1.87)
Pop	0.548 (1.8)	0.092 (0.12)	0.548 (2.01)
recomAccurate	−3.263 (2.01)	−0.208 (0.1)*	−3.262 (1.93)
recomUseful	4.288 (1.86)*	0.331 (0.1)***	4.288 (1.92)*
buyingFreq	2.806 (2.14)	0.108 (0.11)	2.806 (2.07)
songsOwned	−1.819 (2.51)	−0.122 (0.17)	−1.819 (2.46)
Constant	−32.288 (26.94)	−1.592 (1.43)	−32.286 (26.69)
R^2	0.1881	0.2455	
χ^2	128.16***	201.78***	109.06***

Notes. Number of clusters = 72, $n = 1{,}080$ (72 participants × 15 responses, for each analysis). Standard errors are in parentheses, all models use robust standard error estimation, clustered by participants. Model summaries: Model 1— GLS estimation with random participant-level effects; Model 2—log-normal GLS (i.e., dependent variable = $\ln(WTP + 1)$) with random participant-level effects; Model 3—Tobit regression (upper limit 99, lower limit 0) with random participant-level effects. All models estimated using the Stata 14 software.

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

during the study, e.g., due to fatigue, inattention, or a declining lack of trust in the recommender system in response to the recommendations being received. If such order effects exist, we would consequently expect a lowering of the effect of the recommendations on users' judgments over time. By contrast, if the effects of the recommendations are persistent over the series of pricing responses, then no concerns about possible order effects are raised. The relevant test is of the interaction term between the new variable *DisplayOrder* and *ShownRating* (after mean centering the variables). The baseline model utilized GLS regression with random participant-level effects,[12] the same set of control variables, and control for the participants' preferences using the predicted rating for each song in the study. A log-normal GLS regression with random participant-level effects (Model 2) and a Tobit regression with random participant-level effects (Model 3) were also estimated and compared to account for the distribution of WTP. Robust standard errors, clustered by participants, were used in all models. The regression results are presented in Table 7. The control variables that were significantly related to the dependent

variable, *WTP*, show no consistency across the three studies and are not discussed.

The results for Study 3 are consistent with H3 and in line with the ANOVA. There is a significant effect of shown recommendations on consumers' pricing behavior. Even with a lowering of uncertainty due to forced sampling of the priced songs, the randomly generated recommendations significantly affected consumers' willingness to pay. Looking first at the lognormal model (Model 2), we observed an increase (decrease) of 10.5% in willingness to pay for each 1-star increase (decrease) in the shown recommendation rating. The GLS model provides similar results: a 1-star increase (decrease) in the shown recommendation results in a 1.998 cents U.S. increase (decrease) in willingness to pay, in a sample with an average willingness to pay of approximately 20.36 cents U.S. The Tobit model provides similar results, with a 1.999 cents U.S. marginal effect on the uncensored willingness to pay (i.e., unobserved latent variable y^*) and a 1.577 cents U.S. adjusted marginal effect that takes into account censoring. Together, the regression results suggest that we can conservatively expect a positive effect of approximately 7.7%–11.5% in willingness to pay for each 1-star increase in shown rating even when subjects are forced to sample the product prior to pricing.

The display order of songs and its interaction with shown rating had nonsignificant coefficients across all three models (Table 7). The effect of recommendations on preference construction and willingness to pay is consistent throughout the entire experiment. Our presumption of a lack of fatigue and persistent attention are not challenged, and there is no evidence of a decreased effect of the recommendations on participants' willingness to pay, as the participants evaluated a large number of songs during the session.

As an additional post hoc and exploratory analysis, we compared the means of each treatment group between Studies 1 and 3. Recall that the main difference in the design between Studies 1 and 3 is that in Study 3 participants were required to listen to a full 30-second sample prior to making a willingness to pay judgment.[13] Study 2 is not included in this comparison because the design is fundamentally different from Studies 1 and 3. Table 8 presents the resulting *p*-values of the standard *t*-tests with unequal variance, Monte Carlo Fisher–Pitman permutation tests, and the

Wilcoxon rank-sum tests used to compare treatment means. As can be seen, there are no statistically significant differences in means between corresponding treatment groups across the two studies. The results suggest that reduced uncertainty through forced sampling does not appear to reduce the effects of recommendations on preference construction and willingness to pay. Something other than uncertainty in preferences is driving the observed effects in this situation. Many online retail sites offer the capability of sampling prior to purchase. Although sampling reduces the uncertainty of preferences at the point of sale, it does not significantly reduce the effect of system recommendations on preference judgments. This result and the results of Studies 1 and 2 suggest that the information-integration explanation is a plausible candidate as an operative mechanism driving the effects of personalized recommendations on willingness to pay.

6. Discussion and Conclusions

In three laboratory experiments, we examined the impact of recommendations on consumers' economic behavior. The research integrates ideas from behavioral economics and recommender systems, both from practical and theoretical standpoints. The willingness-to-pay studies were performed using judgments from young adults for music, a highly relevant stimulus for which recommendations are readily available in the marketplace. The main contribution of this work is the identification and measurement of a *robust and strong* side effect of system recommendations on consumers' willingness to pay. We observed marginal effects ranging from 7%–17% in willingness to pay for 1-star changes in recommendations. This overarching result suggests that recommender systems in modern electronic commerce are not only decision aids for reducing search costs, but coincidentally may also play a significant role in economic decision making.

Study 1, through a randomized trial design, demonstrated that online recommendations can affect willingness-to-pay judgments, even when the two are measured on different scales. Study 2 extended these results to demonstrate that the same effects exist for real recommendations that contain errors. We introduced systematic errors to recommendations that were

Table 8. Tests of Treatment Means Between Studies 1 and 3

	Study 1	Study 3	P(2-tailed t-tests)	P(2-tailed Fisher–Pitman permutation test)	P(2-tailed Wilcoxon rank-sum test)
Control	24.30 (25.95)	20.62 (25.42)	0.3076	0.3070	0.2878
Low	16.85 (21.38)	17.53 (23.32)	0.8490	0.8686	0.7443
Mid	23.32 (24.04)	21.24 (25.17)	0.4561	0.4538	0.2605
High	26.45 (27.80)	22.93 (26.75)	0.3877	0.3796	0.3719

calculated using state-of-the-art recommendation algorithms used in practice and still observed strong and linear effects. Study 3 forced participants to listen to song samples prior to making their pricing decisions, thereby reducing the uncertainty that has been claimed as a possible factor precipitating the effects of system recommendations. The effects persisted.

From a theoretical perspective, the studies further our understanding of the impact of system recommendations on stated preferences. Recommender systems are an increasingly available decision tool for consumers in online settings that are connected to multibillion-dollar retail businesses. Compared to previous work, our attention shifts from the preference ratings to the willingness-to-pay judgments, in which we introduce several differences from previous research. We demonstrate anchoring-related effects in a realistic preference setting, whereas most prior anchoring research has been directed at judgments in response to general knowledge questions (e.g., Chapman and Johnson 2002; as noted by Ariely et al. 2003). Our studies also extend earlier work that investigated the effects of recommendations on preference ratings (e.g., Cosley et al. 2003, Adomavicius et al. 2013). In these prior studies, there were few consequences and no real economic costs to participants in providing preference ratings. We focused on the real economic impacts of the judgments being made using the incentive compatible BDM procedure; participants' decisions resulted in actual purchases. Furthermore, our controlled laboratory experiment allowed us to use randomization for identifying the causal effects of personalized recommendations, an advantage over prior work that primarily used field studies or secondary data from real-world retailers.

In addition, we provide evidence about the proposed mechanisms underlying the observed effects. Our results indicate that the effect of recommender systems is not attributable purely to a scale compatibility effect. We purposefully designed the current studies using system ratings provided on a 5-star rating scale, while measuring subjects' pricing behavior along a completely different scale, 0–99 cents U.S. Another proposed explanation is based on the uncertainty of preferences, i.e., that in forming a judgment the person searches from the recommendation to the first feasible value in a distribution or range of uncertain values, leading to final estimates tilted toward the recommendation (e.g., Jacowitz and Kahneman 1995). By requiring participants to sample songs prior to making willingness-to-pay judgments, we reduced the role of uncertainty; yet, the observed recommendation effects persisted without abatement. A leading, remaining explanation proposed for the observed effect, therefore, is that recommendations are seen as informative in the formation of preferences, generally, both before purchasing and following consumption.

There are also significant practical implications of the results. The findings raise new issues regarding the design of recommender systems. Arguably, since preferences are subjective, they should be unaffected by recommendations when formed immediately after experiencing the product. Adomavicius et al. (2013) suggested this was not the case for preference ratings, and our research indicates this is also not the case for purchasing judgments. Recommendations are plausibly useful, and perceived so by consumers, when preferences are uncertain; however, our research indicates this information effect extends to rating and willingness-to-pay judgments formed even immediately after the experience, i.e., when uncertainty is low.

Many large online companies rely heavily on recommender systems in their retail practice (e.g., Amazon, iTunes, and Netflix). An effect of the size observed in our studies could have significant impacts on revenues and profitability. More generally, the relationship between the microlevel effects studied here and the macrolevel effects, as well as the relationship between the effects of personalized recommendations and aggregate ratings, both discussed in Section 2.1, are interesting potential areas of further study. The true nature of these impacts on total welfare may not be so obvious. For example, recommendation errors that result in underestimating consumers' true preferences could potentially hurt online retailers, since lower recommendations would pull down consumers' willingness to pay for items. Alternatively, recommendation errors that result in inflated recommendation ratings could erode consumer surplus.

In general, the issue of bias due to recommendations is not trivial and can potentially have a net negative impact on online sales environments and the retail economy. As identified by Cosley et al. (2003), biases in consumer preferences because of recommender systems raise several potential issues, including the following: (1) biases can contaminate the inputs of the recommender system, reducing its effectiveness; (2) biases may provide a distorted view of the system's performance; or (3) biases might allow agents to manipulate the system so that it operates in their favor. A question that arises is: Can we reduce the biases that recommendations introduce while maintaining the benefits that they provide? One approach would be to mechanically adjust the recommendation algorithms used by consumers in judgments to correct the bias. A second approach may be to train consumers about the bias being exhibited, similar to how we teach about the effects of advertising. Yet another approach may be to better design the decision architecture (Thaler and Sunstein 2008) in recommender systems, e.g., by reconstructing the judgment interface.

Consequently, we strongly recommend pursuing further research in this area. As suggestions, we highlight a few areas as potential foci. First, field studies and

analysis of secondary sales data from online retailers could help identify the true magnitude and impact of biases due to recommendations. Second, recommender system designs and implementations should be evaluated for potential impacts on consumer economic behavior from the standpoint of their decision architecture. There is a significant opportunity to rethink the way recommender systems calculate and present recommendations to users by taking into account the effects of system-induced biases. Third, the evaluation of recommender systems may need to be reengineered to consider the potential for biases in the consumer preference input. If the inputs to the system are biased, simple comparisons of generated recommendations to reported preferences may no longer be the most effective means of measuring system performance. Fourth, our study was limited to a within-session time horizon. In other words, we consistently observe short-time recommendation effects on consumers' willingness to pay. It is worthwhile to explore if the observed effects persist across multiple sessions and over longer time periods. Fifth, although less central, our study provides some insights into the mechanisms that may or may not underlie the effects of system recommendations on individual preferences; additional investigations into these mechanisms would be useful. Finally, further research is needed to investigate how the effects of personalized recommendations compare to the effects of nonpersonalized, general item quality information (e.g., online aggregate user ratings, user reviews, best seller lists). Are the effects of personalized and nonpersonalized item quality information connected? How do these two types of information operate in conjunction? We look forward to ongoing investigations.

Acknowledgments
The authors thank the senior editor, the associate editor, and the anonymous review team for their thoughtful guidance during the review process.

Endnotes
[1] In this paper, we focus on traditional application domains of recommender systems, i.e., domains of experience goods where consumer preferences are horizontal or *taste-driven* (Chen and Xie 2005). This is in contrast to domains where consumer preferences are vertical, i.e., where there is a more objective understanding about product quality. In the latter domains, personalized recommendations to individual users are much less relevant (e.g., as compared to review systems) since consumers tend to have shared preferences based on quality.

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

[3] http://www.billboard.com/articles/news/7400248/pwc-global-entertainment-study.

[4] As a check, we performed the analyses in all three studies including the omitted subjects as well. None of the results were affected by removing these inappropriately recruited subjects.

[5] Billboard Year-End Hot 100 Songs. http://www.billboard.com/charts/year-end/2006/hot-100-songs.

[6] We were generally successful in offering sufficient numbers of songs that were judged favorably. For the ratings received in Task 1 described in Section 3.1.4, 48% of the songs were rated positively. The willingness to pay for songs, as shown in Figure 2, also supports the sufficiency of our song database.

[7] The study had four tasks, which will be described in Section 3.1.4.

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

[9] For two participants, 40 nonowned songs were not able to be identified: one subject identified only 24 songs and the other identified only 16 songs as not-owned. The removal of these two subjects did not change the results, so they were included in all relevant analyses. In the third task, one subject saw 10 songs in each of the *Low* and *Mid* conditions, 4 in the *High* condition, and no *Control* condition songs. One subject saw 10 songs in the *Low* condition, 6 in the *Mid* conditions, and no *High* or *Control* condition songs.

[10] We also tested models with fixed participant-level effects in both studies and observed the same results in terms of significance, direction, and approximate magnitude. Online Appendix 3 reports fixed-effect analysis results. We also tested for nonlinear relationships but no significant curvature was indicated.

[11] We also compared the control conditions across all three studies reported in the paper. Overall, the difference in willingness to pay across all three groups of participants was only marginally significant ($F(2, 1036) = 2.96$, $p = 0.052$), with the mean responses being somewhat larger in Study 1 compared to Studies 2 and 3. However, these slight possible differences between subject groups do not impact the key comparisons, which are all within-subjects comparisons.

[12] As with Studies 1 and 2, models using fixed-effects and bootstrapped errors for Study 3 are presented in Online Appendix 3. The results are consistent with those presented in Table 7.

[13] Note that Studies 1 and 3 occurred approximately two years apart but used the same set of songs in the experiment. In Study 1, multiple song recommendations were presented on a single page, whereas in Study 3, only one song is presented per page. This analysis is purely exploratory and post hoc. However, we emphasize that it can only strengthen our main finding that the effects of recommendations on willingness to pay do not subside when preference uncertainty is significantly reduced.

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

Adomavicius G, Bockstedt J, Curley S, Zhang J (2013) Do recommender systems manipulate consumer preferences? A study of anchoring effects. *Inform. Systems Res.* 24(4):956–975.

Amatriain X, Basilico J (2012) Netflix recommendations: Beyond the 5 stars (part 1). *Netflix Tech Blog* (April 5). http://techblog.netflix.com/2012/04/netflix-recommendations-beyond-5-stars.html.

Ariely D, Loewenstein G, Prelec D (2003) "Coherent arbitrariness": Stable demand curves without stable preferences. *Quart. J. Econom.* 118(1):73–105.

Ariely D, Loewenstein G, Prelec D (2006) Tom Sawyer and the construction of value. *J. Econom. Behav. Organ.* 60(1):1–10.

Becker GM, DeGroot MH, Marschak J (1964) Measuring utility by a single-response sequential method. *Behav. Sci.* 9(3):226–232.

Bennet J, Lanning S (2007) The Netflix Prize. *KDD Cup and Workshop* (ACM, New York), 3–6.

Box GEP (1954a) Some theorems on quadratic forms applied in the study of analysis of variance problems, I. Effect of inequality of variance in the one-way classification. *Ann. Math. Statist.* 25(2):290–302.

Box GEP (1954b) Some theorems on quadratic forms applied in the study of analysis of variance problems, II. Effects of inequality of variance and of correlation between errors in the two-way classification. *Ann. Math. Statist.* 25(3):484–498.

Camerer C, Hogarth R (1999) The effects of financial incentives in experiments: A review and capital-labor-production framework. *J. Risk Uncertainty* 19(1/3):7–42.

Chapman G, Bornstein B (1996) The more you ask for, the more you get: Anchoring in personal injury verdicts. *Appl. Cognitive Psych.* 10(6):519–540.

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

Chen Y, Xie J (2005) Third-party product review and firm marketing strategy. *Marketing Sci.* 24(2):218–240.

Cosley D, Lam S, Albert I, Konstan JA, Riedl J (2003) Is seeing believing? How recommender interfaces affect users' opinions. *CHI 2003 Conf.* (ACM, New York), 585–592.

Donaldson C, Jones A, Mapp T, Olson JA (1998) Limited dependent variables in willingness to pay studies: Applications in health care. *Appl. Econom.* 30(5):667–677.

Epley N, Gilovich T (2010) Anchoring unbound. *J. Consumer Psych.* 20(1):20–24.

Fleder D, Hosanagar K (2009) Blockbuster culture's next rise or fall: The impact of recommender systems on sales diversity. *Management Sci.* 55(5):697–712.

Godinho de Matos M, Ferreira P, Smith MD, Telang R (2016) Culling the herd: Using real-world randomized experiments to measure social bias with known costly goods. *Management Sci.* 62(9):2563–2580.

Hosanagar K, Fleder D, Lee D, Buja A (2014) Will the global village fracture into tribes? Recommender systems and their effects on consumer fragmentation. *Management Sci.* 60(4):805–823.

Jacowitz KE, Kahneman D (1995) Measures of anchoring in estimation tasks. *Personality Soc. Psych. Bull.* 21(11):1161–1166.

Kahneman D (2011) *Thinking, Fast and Slow* (Farrar, Straus and Giroux, New York).

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

Marshall M (2006) Aggregate knowledge raises $5 m from Kleiner, on a roll. *Venture Beat* (December 10). Accessed April 29, 2012, http://venturebeat.com/2006/12/10/aggregate-knowledge-raises-5m-from-kleiner-on-a-roll/.

Mauchly JW (1940) Significance test for sphericity of a normal *N*-variate distribution. *Ann. Math. Statist.* 11(2):204–209.

Müller H, Kroll EB, Vogt B (2012) Do real payments really matter? A re-examination of the compromise effect in hypothetical and binding choice settings. *Marketing Lett.* 23(1):73–92.

Mussweiler T, Strack F (1999) Hypothesis-consistent testing and semantic priming in the anchoring paradigm: A selective accessibility model. *J. Experiment. Soc. Psych.* 35(2/3):136–164.

Northcraft G, Neale M (1987) Experts, amateurs, and real estate: An anchoring-and-adjustment perspective on property pricing decisions. *Organ. Behav. Human Decision Processes* 39(1):84–97.

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

Salganik MJ, Watts DJ (2008) Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Soc. Psych. Quart.* 74(4):338–355.

Salganik MJ, Dodds PS, Watts DJ (2006) Experimental study of inequality and unpredictability in an artificial cultural market. *Science* 311(5762):854–856.

Sarwar B, Karypis G, Konstan JA, Riedl J (2001) Item-based collaborative filtering recommendation algorithms. *10th Internat. WWW Conf.* (ACM, New York), 285–295.

Schkade DA, Johnson EJ (1989) Cognitive processes in preference reversals. *Organ. Behav. Human Decision Processes* 44(2):203–231.

Shapiro C, Varian HR (1999) *Information Rules: A Strategic Guide to the Network Economy* (Harvard Business Review Press, Boston).

Surowiecki J (2004) *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies, and Nations*, 1st ed. (Doubleday, New York).

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

Tucker C, Zhang J (2011) How does popularity information affect choices? A field experiment. *Management Sci.* 57(5):828–842.

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases. *Science* 185(4157):1124–1131.

Tversky A, Sattath S, Slovic P (1988) Contingent weighting in judgement and choice. *Psych. Rev.* 95(3):371–384.

Wilson TD, Houston CE, Etling KM, Brekke N (1996) A new look at anchoring effects: Basic anchoring and its antecedents. *J. Experiment. Psych.* 125(4):387–402.

Wright WF, Anderson U (1989) Effects of situation familiarity and financial incentives on use of the anchoring and adjustment heuristic for probability assessment. *Organ. Behav. Human Decision Processes* 44(1):68–82.

Zhang J, Liu P (2012) Rational herding in microloan markets. *Management Sci.* 58(5):892–912.