Self-Selection and Information Role of Online Product Reviews

Xinxin Li, Lorin M. Hitt,

Please scroll down for article—it is on subsequent pages

Self-Selection and Information Role of Online Product Reviews

Xinxin Li
School of Business, University of Connecticut, Storrs, Connecticut 06269,
xli@business.uconn.edu

Lorin M. Hitt
The Wharton School, University of Pennsylvania, Philadelphia, Pennsylvania 19104,
lhitt@wharton.upenn.edu

Online product reviews may be subject to self-selection biases that impact consumer purchase behavior, online ratings' time series, and consumer surplus. This occurs if early buyers hold different preferences than do later consumers about the quality of a given product. Readers of early product reviews may not successfully correct for these preference differences when interpreting ratings and making purchases. In this study, we develop a model that examines how idiosyncratic preferences of early buyers can affect long-term consumer purchase behavior as well as the social welfare created by review systems. Our model provides an explanation for the structure of product ratings over time, which we empirically test using online book reviews posted on Amazon.com. Our analysis suggests that firms could benefit from altering their marketing strategies such as pricing, advertising, or product design to encourage consumers likely to yield positive reports to self-select into the market early and generate positive word-of-mouth for new products. On the other hand, self-selection bias, if not corrected, decreases consumer surplus.

Key words: online product reviews; self-selection; consumer heterogeneity; herding
History: Anil Gupta, Senior Editor. This paper was received on June 13, 2006, and was with the authors $4\frac{1}{2}$ months for 2 revisions. Published online in *Articles in Advance* July 31, 2008.

1. Introduction

Word-of-mouth has long been recognized as a major driver of product sales. Not only can word-of-mouth generally increase consumer awareness, but it may also be one of the only reliable sources of information about the quality of experience goods (i.e., products not easily characterized prior to consumption). With the development of the Internet, word-of-mouth has moved beyond small groups and communities to being freely available through large-scale consumer networks (Avery et al. 1999). These networks have magnified the depth and span of word-of-mouth to an unprecedented scale. Online opinion and consumer-review sites have correspondingly changed the way consumers shop, enhancing or even supplanting traditional sources of consumer information such as advertising. In a survey of 5,500 Web consumers conducted by BizRate, 44% of respondents said they had consulted opinion sites before making a purchase and

59% considered consumer-generated reviews more valuable than expert reviews (Piller 1999). In some product categories such as electronics, surveys suggest that online review sites have a greater influence on purchase decisions than any other medium (DoubleClick 2004).

A large body of work has analyzed the design and performance of eBay-like online reputation systems (see a survey in Dellarocas 2003). However, there has been considerably less systematic research on product review websites. Several studies have shown a link between online reviews and product sales (Godes and Mayzlin 2004, Chevalier and Mayzlin 2006). However, these studies did not directly address whether online reviews effectively communicate information about quality. There are at least two reasons why online reviews may fail to provide information about quality. First, firms may manipulate reviews to create artificially high ratings (such as by using paid reviewers),

although theoretical results by Dellarocas (2006) suggest that manipulated reviews are still informative. Second, even if reviews accurately reflect earlier consumers' opinions, those opinions may not be representative of the opinions of the broader consumer population in later time periods. In particular, for goods that have elements of both vertical and horizontal differentiation, ratings may represent a mix of objective product quality and subjective assessments of value based on consumer fit. If the preferences of a product's early adopters—adopters who also post the first reviews—systematically differ from the broader consumer population, the early reviews can be biased. This bias is a "self-selection bias" because products are not randomly assigned to reviewers. Rather, early buyers self-select products that they believe they may enjoy. For example, new releases of books are often purchased by avid fans of the authors' previous books; they may tend to assign higher ratings than do consumers in the general population. The existence of positive, self-selected early-review bias may explain why reviews of most products tend to fall over time.[1] Alternatively, early adopters in some categories may be more sensitive to advanced "cutting-edge" features in the product, which may cause their perceptions of product quality to be different from those of the general population who more equally weight all features or who prefer other features such as "ease-of-use" or "simplicity." This type of bias could be either positive or negative.

If consumers can correct for differences in reviewer taste when considering their purchases, these biases may not affect sales. However, discerning and correcting for reviewers' preference function may be difficult, especially if consumers solely rely on a numeric scale (e.g., the Amazon "star rating system") in making purchases. Thus, self-selection by early buyers can create bias in reviews which affects sales, even if all reviews are truthful. In this paper, we explore the presence and implications of this self-selection bias by addressing several research questions:

• Does self-selection bias exist in online reviews?

• Do consumers correct for this bias when making purchase decisions?

• How does review bias affect market outcomes (sales and consumer surplus)?

• How should firms adjust their strategies to account for self-selection review bias?

We begin by constructing a theoretical model of buyers' self-selection behavior that explains the patterns in reviews over time as well as enables the analysis of consumer welfare and firm profits under the existence of self-selection bias. Next, we formulate hypotheses and empirically evaluate the assumptions underlying our theoretical model by first analyzing a time series of rating averages for a large sample of books sold on Amazon.com. Next, we decompose the numerical ratings from individual reviews into a component related to self-selection and a component related to overall quality, and examine whether either or both components are correlated with sales. Finally, we discuss the impacts of this self-selection behavior on consumer welfare and firm strategies.

2. Literature Review

Even before the emergence of large-scale online communication networks, word-of-mouth was perceived as an important driver of product sales (Rogers 1962; see the summary in Lilien et al. 1992, Chapter 10). Most of these studies focused on the diffusion of positive experience, which is more related to raising consumer awareness than it is to conveying quality information.[2] In addition, most of this work focused on relatively small communities. The emergence of large-scale online communication networks for the exchange of quality information has led to an emerging literature on the economics of these systems. Considerable research has focused on performance and design of eBay-like reputation systems (see a comprehensive review by Dellarocas 2003). However, research on product review systems has been more limited. Chevalier and Mayzlin (2006) demonstrated that the differences between consumer reviews posted on Barnes & Noble and those posted on Amazon.com were positively related to the differences in book sales via the two websites. Godes and Mayzlin (2004) showed in a different setting that the "dispersion" of conversations

---

[1] The trend of the review ratings would be flat if consumers' preferences over product attributes were identical or if preferences across consumers differ but early buyers purchase and review randomly.

[2] An exception is Mahajan et al. (1984), who incorporated negative word-of-mouth in a model of advertising timing.

about TV shows across online consumer communities and the popularity of these TV shows were strongly related. Dellarocas et al. (2004) found that the valence (average numerical rating) of online consumer reviews is a better predictor of future movie revenues than other measures they considered. In contrast, Duan et al. (2005) proposed the importance of the number of online reviews in influencing box office sales. Clemons et al. (2006) found that the variance of ratings and the strength of the most positive quartile of reviews have a significant impact on the growth of craft beers. Chen and Wu (2004) suggested the mediation role of product recommendations in affecting the relationship between reviews and sales on Amazon.com. Although these studies have established a link between sales and product reviews, they did not examine whether consumer reviews were effective in communicating actual product quality. Moreover, these studies utilized the time-series dimension of the data to increase the sample size but did not directly address the time structure of reviews.

As mentioned earlier, one reason why consumer-generated reviews may not represent actual product quality is due to "forum manipulation," in which firms hire professional reviewers (or encourage friends and colleagues) to artificially boost the ratings of their products. In a recent paper, Dellarocas (2006) discussed how this occurs in practice and presented a theoretical model for producers' optimal investment in forum manipulation. His results suggest that even in the presence of manipulation, reviews are still (or even more) informative because producers of the highest-quality products also receive the greatest benefit from manipulation.

In our analysis, we focus on a setting in which reviews are truthful but may be misleading due to differences in preferences between earlier product buyers (and therefore early reviewers) and later product purchasers. This explanation appears to be new in the literature. However, this scenario only becomes interesting if later consumers do not account for this early-reviewer bias when making purchase decisions. The observation that people tend to follow the decisions of others has been extensively discussed in the herding literature, which has attributed this behavior to network externalities (Katz and Shapiro 1985), social

sanctioning of deviants (Akerlof 1980), and taste for conformity (Becker 1991). Our work is more closely related to information-motivated herding literature (Banerjee 1992, Bikhchandani et al. 1992) because it is the quality information indicated by early buyers' reviews or ratings that drives later buyers to follow. However, in the cited models, buyers share similar quality perceptions, so herding is the result of rational behavior. In our paper, in contrast, buyers differ in preferences, so following the advice provided by early biased reviews is not necessarily optimal. This behavior of suboptimal following is consistent with bounded rationality (Kahneman 2003, Rabin 1998) in the sense that an individual may follow simple decision rules that lead to suboptimal choices in complicated contexts. This behavior is also supported by empirical results in online auctions that suggest that bidders prefer auctions with more bidders even when more bidding is not indicative of product quality (Simonsohn and Ariely 2005).

3. Theoretical Analysis

3.1. The Model
Consider a market for an experience good in which, in each period, a group of consumers comes into the market and makes a decision about whether to purchase (at most) one unit of the product. We consider the scenario in which the good is a durable good and there is no repeat purchase. We believe that consumer reviews have the most impact on this type of product because quality cannot be revealed simply by consumer testing of all alternatives.

Consider a product that has two sets of attributes. One set of attributes can be inspected before purchase and the other cannot. For example, the author of a book can be inspected before purchase but the content cannot; the cast of a movie can be inspected before purchase but the script cannot; the brand of a skin care product can be inspected before purchase but the effectiveness cannot. We define the set of attributes that can be inspected before purchase to be "search attributes" and the set of attributes that cannot be inspected before purchase to be "experience attributes" or "quality." An individual consumer's preferences over the product can be characterized

by two components (x_i, q_i).[3] The value of element x_i represents the preferences of consumer i over the "search attributes" of the product and is known by each consumer before purchasing. The value of element q_i measures the "quality" of the product for consumer i—each consumer may perceive quality of the same product differently. Consumers only learn q_i after buying the product. The net utility of consuming the product for consumer i is defined as $U(x_i, q_i, p) = \tau \cdot q_i + x_i - p$, in which p is the price of the product which is assumed fixed across time. The parameter τ determines the relative importance of postpurchase experience-related attributes versus prepurchase "search attributes" in consumers' valuations of the product.

Assume that x_i is uniformly distributed over $[0, 1]$ (denote the mean as u_x and variance as σ_x^2) and q_i satisfies a symmetric beta distribution with parameter s (denote the mean as q and variance as σ_q^2).[4] We choose the symmetric beta distribution for q_i for two reasons. First, this distribution is bounded between zero and one; second, as the value of s varies, the magnitude of consumer heterogeneity changes without expanding the range of values for q_i. Before buying and experiencing the product, consumers have a prior expectation over q, denoted as q^e, and this expectation can be affected by published online consumer reviews. Without loss of generality, assume that q^e is the same for all consumers.[5] If there are no product reviews in the market, q^e is arbitrarily picked by consumers, which is assumed to be a low initial value often associated with unfamiliar products—denoted as q^e_{prior}.[6] Consumers will make their purchase decisions based on their expected utility.

Expanding on the previous literature, we allow x_i and q_i to be correlated with correlation coefficient ρ. Then, given x_i, the expected value of q_i (denoted as

$E[q_i \mid x_i]$) can be approximated by $q + \rho \cdot \sigma_q(x_i - u_x)/\sigma_x$. The parameter ρ describes the correlation between demand and quality perception. That is, consumers who are more likely to be attracted by the "search attributes" of the product and consequently buy early may be more likely to think of the product as high quality compared to the consumer population. To illustrate, consider the book market we examine in this paper's empirical study. The readers who tend to buy early are probably fans of the author's previous books and, thus, are more likely to enjoy the book. A similar situation is also true for consumers of game software—early buyers tend to be the most enthusiastic supporters of new games and tend to be tolerant of early-stage "beta" software defects. These examples correspond to an instance in which x_i and q_i are positively correlated. In contrast, early adopters of some types of commercial software may be very sensitive to software defects. This would correspond to a negative correlation between x_i and q_i.

In each period, consumers who bought the product will post their (truthful) product evaluations online for access by all future buyers. Because consumers may perceive quality of the same product differently and because their reviews reflect their respective personal tastes, whether these product reviews can communicate the actual average quality (q) depends on whether the consumers who post reviews are representative of the population. If x_i and q_i are correlated, then the consumers who tend to buy early and thus write reviews early are systematically biased, which in turn affects the demand for the product and the types of consumers that purchase the product in future periods.

In the first period, there are no product reviews available in the market, so $q_1^e = q_{prior}^e$. Without loss of generality, we normalize the value of the best alternative to this product to be zero. Thus, only consumers with expected utility $U(x_i, q_1^e, p)$ larger than zero will buy the product. First-period demand equals $1 - \text{Min}\{1, \text{Max}\{0, p - \tau \cdot q_1^e\}\}$. Unless $p - \tau \cdot q_1^e \geq 1$ (no one buys in the first period), the consumers who bought the product will post quality evaluations online at the end of the first period, and the average rating $\bar{R}_1$ has an expected value of $q + \rho \cdot \sigma_q((1 + \text{Max}\{0, p - \tau \cdot q_1^e\})/2 - u_x)/\sigma_x$. The review

---

[3] Similar assumptions are used in Villas-Boas (2004) to describe a consumer's preferences over observed features of the products and product quality.

[4] The probability density function is $w(q_i) = (1 - q_i)^{s-1} q_i^{s-1}/Beta(s, s)$.

[5] If q^e differs across consumers, then we can include q_i^e in x_i and the subsequent analysis still applies.

[6] Prior expectation is also assumed as exogenously given in Shapiro (1983a) and Schmalensee (1982). Shapiro (1983b) also points out that consumers' expectations about new product quality are generally not fully rational.

bias, the average difference between $E[\bar{R}_1]$ (the average quality rating posted online) and q (the actual average quality), is zero only if the buyers in the first period are representative of the general population. That is, reviews will have self-selection bias unless there is no correlation between x_i and q_i (ρ is zero) or unless all consumers who arrived in the first period purchased the product ($p - \tau \cdot q_1^e < 0$). If the consumers who tend to buy early are the ones who are more likely to appreciate the product's experience (ρ is positive), there is a positive self-selection bias reflected in the reviews and the bias increases with the correlation between demand and quality perception (ρ), the degree of consumer heterogeneity on quality perception (σ_q), and the deviation of early buyers from an average consumer ($((1+\text{Max}\{0, p - \tau \cdot q_1^e\})/2 - u_x)/\sigma_x$). If ρ is negative, the review bias is negative; its magnitude increases with ρ and decreases with σ_q ($((1+\text{Max}\{0, p - \tau \cdot q_1^e\})/2 - u_x)/\sigma_x$).

In each subsequent period, a new group of buyers comes into the market and reads the reviews posted by those who bought the product in all preceding periods. We assume, for our model, that consumers take reviews as representative of their preferences and make purchase decisions accordingly (even if reviews are biased). This assumption can be justified by the cost and difficulty for consumers to infer their valuation of a product based on reviews of other consumers. When the information relevant to discerning the reviewers' characteristics is limited or the process of evaluating reviews is costly, consumers may simply take the average rating as indicative of their preferences. This suboptimal market following behavior is consistent with bounded rationality (Kahneman 2003, Rabin 1998) as discussed in §2, and can be justified by empirical evidence in §4.

Given this model setup, in the kth period a new consumer believes that the reported reviews accurately estimate the average quality level she is going to experience (in other words, she perceives ρ to be zero regardless of the true value). She therefore updates her belief on quality as equal to $\bar{R}_{k-1}$, the average rating of all reported reviews posted before the kth period. Consumers with $U[x_i, \bar{R}_{k-1}, p]$ greater than zero will buy the product and post their reviews

by the end of this period. In the end of period k, the expected value of the average rating is

$$E[\bar{R}_k] = \begin{cases} E[\bar{R}_{k-1}] & \text{if } p - \tau \cdot \bar{R}_{k-1} \geq 1, \\ q + \rho \cdot \dfrac{\sigma_q}{\sigma_x}\bigg[\bigg(\sum_{n=1}^{k-1}(1 - \text{Max}\{0, p - \tau \cdot \bar{R}_{k-1}\}) \\ \qquad\qquad \cdot (1 + \text{Max}\{0, p - \tau \cdot \bar{R}_{k-1}\})\bigg) \\ \quad \cdot \bigg(2\sum_{n=1}^{k-1}(1 - \text{Max}\{0, p - \tau \cdot \bar{R}_{k-1}\})\bigg)^{-1} - u_x\bigg] \\ \qquad\qquad \text{if } p - \tau \cdot \bar{R}_{k-1} < 1. \end{cases}$$

This will further affect the demand in the $(k + 1)$th period.

Unlike consumers, we assume that firms know the true value of ρ, which can presumably be determined through market research such as consumer surveys, focus groups, or analysis of purchase data of similar products. The firm chooses the optimal price based on its knowledge of ρ to maximize total profit $p \sum_k (1 - \text{Min}\{1, \text{Max}\{0, p - \tau \cdot \bar{R}_{k-1}\}\})$. The product is in market for K periods. We normalize marginal cost to be zero. All model predictions are obtained numerically.

3.2. Hypotheses

Before we proceed to analyze the implication of this self-selection behavior for firm profit and consumer welfare (in §5), we first examine if the assumptions underlying our theoretical model can be evaluated using empirical data:

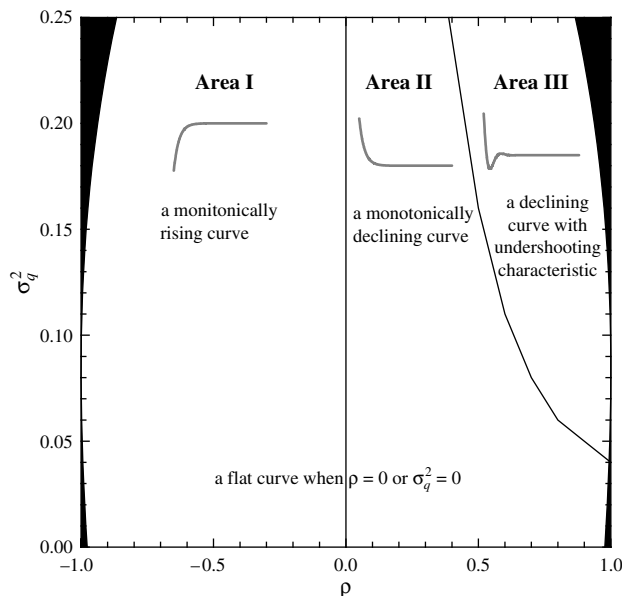(1) Self-selection bias exists in early reviews.

(2) Consumers do not fully correct for the bias when making purchase decisions.

Figure 1 summarizes the time series of average ratings for all possible values of correlation (ρ) and consumer heterogeneity (σ_q^2)[7] under conditions of optimal pricing.[8] This figure can help us validate our assumptions by observing the patterns of reviews in actual data. If demand (determined by x_i) and

---

[7] According to the distribution of q_i, σ_q^2 is between 0 and 0.25. Following Whitt (1976), we calculate the maximum and minimum correlations for each value of σ_q^2 and exclude the correlation values outside of the feasible range (as shown in the dark area in Figure 1).

[8] Trends in reviews will be similar under nonoptimal pricing conditions as long as prices are fixed over time.

Figure 1   **The Patterns in Reviews for All Possible Values of Correlation (ρ) and Variance (σ_q^2)**

quality perception (q_i) are independent ($\rho = 0$) or consumers are homogeneous ($\sigma_q^2 = 0$), then the consumers who write reviews are representative of the general population and the average ratings remain constant over time. If correlation is positive, average reviews have positive bias in early periods due to consumer self-selection and the bias declines over time (Area II). The opposite is true for negative correlations in Area I (with negative bias, reviews increase toward the mean over time). These observations lead to our primary hypotheses on the patterns of reviews over time to evaluate our first assumption that self-selection bias exists. If there is no self-selection ($\rho = 0$), reviews should exhibit no time trend whatsoever. However, if early adopters do have different preferences ($\rho \neq 0$), we should observe either a declining or rising trend over time, contingent on the sign of correlation (ρ).

HYPOTHESIS 1A (H1A) (NO SELF-SELECTION, $\rho = 0$). *Reviews exhibit no time trend.*

HYPOTHESIS 1B (H1B) (POSITIVE SELF-SELECTION, $\rho > 0$). *Reviews follow a declining trend over time.*

HYPOTHESIS 1C (H1C) (NEGATIVE SELF-SELECTION, $\rho < 0$). *Reviews follow a rising trend over time.*

These hypotheses (H1A–H1C) represent a mutually exclusive set of possibilities for temporal patterns in review data and can be distinguished by examining reviews over time both in aggregate and for individual products. Indeed, in a population of products, it is possible that any of them could be true, although we expect the condition described in H1B to be the most likely in book review data that we examine in §4. Ideally, we would also like to examine how reviews vary with the correlation parameter. However, because we are not able to measure correlation directly in the empirical data, alternatively, we instead construct a proxy to estimate relative correlation across products. This leads to the following hypothesis:

HYPOTHESIS 2. *The higher the correlation coefficient (ρ) is, the more likely reviews follow a declining trend over time.*

Our second assumption that consumers respond to review bias can be evaluated by examining the existence of undershooting dynamic and the direct relationship between sales and review bias. According to Figure 1, if both the correlation and the consumer heterogeneity are high (Area III), not only do reviews follow a declining trend over time but a time period also appears in which the reviews undershoot the long-run average before recovering. This distinct undershooting dynamic reflects observable purchase mistakes—some consumers who would not have purchased if they corrected for the bias were induced to buy due to the positive review bias, and their substantive disappointment drives the average rating down to a level even lower than the long-term average. If consumers fully corrected for the review bias, we would still observe monotonically rising or declining curves (because the very first reviews would still be biased), but the undershooting pattern in ratings would never appear because consumers would not make purchase mistakes. Therefore, as long as undershooting is observed in actual data, even if for a fraction of products (consistent with Figure 1), it validates our second assumption.[9] The actual proportion is determined by how samples are distributed

---

[9] In reality, other factors, possibly correlated with quality perception, may strengthen or weaken the trend in reviews. Our model can be extended to incorporate those factors by either including them into x_i or creating new variables similar to x_i and correlated

in the two-dimensional parameter space in Figure 1. In addition, according to Figure 1, if review bias is not fully corrected, then undershooting is more likely to appear for products with higher likelihoods of idiosyncratic consumer preferences (σ_q is high) and self-selection bias (ρ is high). Consequently, we consider the following two hypotheses to evaluate our assumption that consumers do not fully correct for review bias:[10]

Hypothesis 3A (H3A). *For at least a fraction of products, reviews follow a declining trend over time with an undershooting period.*

Hypothesis 3B (H3B). *Undershooting is more likely to appear for the products with more heterogeneous consumer preferences and higher self-selection bias.*

A challenge of this analysis, however, is that the bias must be relatively strong to generate a visible undershooting period in the empirical data. Thus, the test of H3A and H3B may be plagued by low statistical power. Therefore, rather than solely focusing on time trends, we also test our second assumption using the well-established relationship between reviews and sales. In particular, if consumers fully correct for review bias, then the sales of a product (at any point in the product life cycle) should be determined by overall long-term quality ratings but not by short-term deviations of the rating with respect to its long-term value. In other words, the sales and ratings time-series data let us test whether consumers as a population are able to discount early reviews and to infer the true population quality estimate even in the presence of bias. The testable implication of this analysis is that if review bias is present and corrected, the sales of a product should be correlated with long-term average reviews but not with the time-varying component of reviews. However, if bias is present and not fully corrected, sales will not only be correlated with long-term average reviews but also with the time-varying component of reviews. If we are able to separate reviews into a temporal

component as well as a long-term average, we can test:

Hypothesis 4A (H4A) (Reviews Affect Sales). *Sales of a product, all else equal, are correlated with the long-term average review.*

Hypothesis 4B (H4B) (Review Bias Affects Sales). *Sales of a product, all else equal (including long-term average review), are correlated with the temporal component of reviews.*

H4A is, in essence, the hypothesis that has been previously examined in the literature on how reviews influence sales, although these tests were done under the implicit assumption that there was no review bias. A finding of support of H4B, however, will further validate our assumption that review bias is not fully corrected.

4. Empirical Analysis: Data and Results

4.1. Data Collection

A sample of 2,651 hardback books was collected from *Books in Print* and the bestseller lists in *Publisher's Weekly*, covering books published from January 2000 to February 2004 that also had a sufficient number of consumer-written book reviews posted on Amazon.com. The sampling methodology is detailed in the appendix.[11] To prevent our analysis from being affected by large price changes and differences in the packaging of the product, we limited our data set to hardback books and focused only on the period between initial introduction and the release of the paperback version.

Using software agents, we collected books' characteristic information from Amazon.com, including ISBN, author(s), publication date, Amazon.com category (e.g., business and investing), and publication dates for corresponding paperback editions.

---

with q_i. Incorporating more factors may better explain the existence of early review bias but not undershooting. Undershooting appears only when consumers do not fully correct for the review bias.

[10] It should be noted that our model does not suggest a comparable period of "overshooting" due to negative correlation.

[11] This sampling method is similar to that chosen in previous studies on online reviews and sales (e.g., Chevalier and Mayzlin 2006). As in previous work, our sampling methodology is intended to represent sales in the market for products that are likely to be influenced by online reviews by including both high-sales books (i.e., those drawn from bestseller lists) and a sample of other books (i.e., those drawn from the broader "in-print" list).

We also collected all consumer reviews posted on Amazon.com from book release until the earlier of either the paperback release date or the end of our data collection period (July 2, 2004). In total, this sampling methodology provided 136,802 single review observations for 2,651 books. Because we use the observations of average reviews to study the patterns over time, we exclude the observations if the average review is based on less than three reviews (because the average is not meaningful in this case) or if the book has less than five observations in total (because a trend can be barely determined in this case). Thus, the number of raw review observations used in the sample is reduced to 135,485 and the number of books included is 2,203. We then aggregate these data to represent a per-day average for all our calculations, which yields a total of 82,131 observations for average reviews given that more than one review can appear on a particular day (we refer to this sample as the *review data set*). The books in our sample had an average of 61 reviews and an average rating of 4.02 (on a 1 to 5 scale), which is similar to the population means reported by Chevalier and Mayzlin (2006) who utilize similar data. Table 1 provides the summary statistics for this sample.

Every Friday during the 18 weeks from March 5 to July 2 in 2004, we collected sales-related data for each book in our review data set from Amazon.com, including sales rank (as a measure of book sales quantity), price, and shipping availability (e.g., "usually ships within 24 hours"). Although we cannot observe sales directly, we can utilize the relationship between sales rank and actual sales found by Schnapp and Allwine (2001) to estimate the sales during the relevant period.[12] Schnapp and Allwine fit the following model to data on sales versus sales rank in 2001: $\mathrm{Log}[Sales] = 9.61 - 0.78\mathrm{Log}[SalesRank]$. We use this model to calculate sales as a function of rank, scaling these estimates to reflect the 43% growth in Amazon.com's book sales from 2001 to 2004.[13] To control for competition from outside retailers and promotions, each Friday we also collected new book prices

---

[12] A similar method is used in Chevalier and Mayzlin (2006). If sales rank instead of sales is used in our analysis, all results hold.

[13] The growth number is calculated based on Amazon.com's 2004 10-K Report and 2005 8-K Report.

Table 1   Summary Statistics for Number of Reviews and Average Ratings in Our Review Data Set

	Number of reviews per book	Average ratings per book
Mean	61.50	4.02
Standard deviation	189.85	0.60
Percentile (%)		
1	7	2.35
5	10	2.90
10	12	3.20
25	18	3.67
50	29	4.10
75	52	4.47
90	103	4.73
95	178	4.84
99	550	5

listed on a price comparison engine (Pricescan.com) for each book in our sample. We chose Pricescan.com for this study because of its popularity as a search engine for book prices and the fact that it returns more results for book prices than most other search engines. Because some of these data may be missing, our sales data consists of 20,473 "book × week" observations for 1,290 books (we refer to these data as the *sales data set*). Table 2 summarizes the descriptive statistics for the four primary variables.
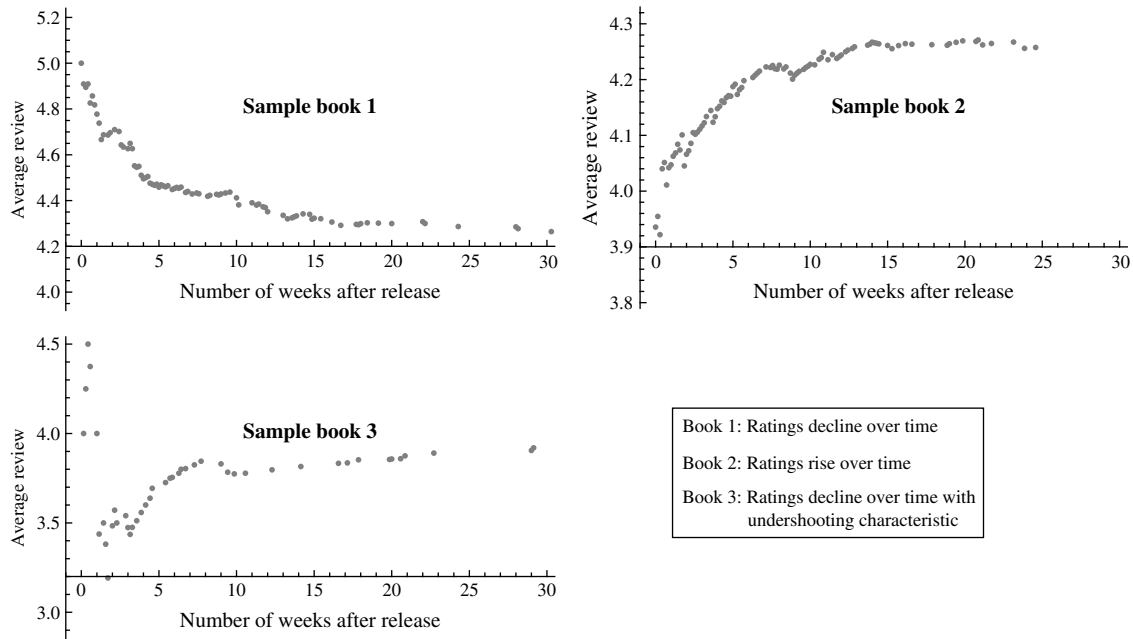
4.2.   Results: Trends in Consumer Reviews

Figure 2 shows the time series of review ratings for three sample books. The figure illustrates some of the dynamics found in review data for the books in our sample. First, reviews are clearly not random but show visually discernable patterns in time. Second, declining and rising trends are both observed, suggesting the presence of both positive and negative biases. Third, some books follow a trend that is almost entirely downward sloping (e.g., sample book 1 in Figure 2), whereas others show a transient period after introduction in which the reviews are considerably lower than

Table 2   Descriptive Statistics for Sales-Related Data in Our Sales Data Set

Variables	Number of observations	Mean	Standard deviation	Minimum	Maximum
Sales rank	20,473	34,527	67,071	3	1,435,765
Price	20,473	17.26	4.71	4.38	81.95
Average rating	20,473	4.07	0.62	1.65	5
Number of reviews	20,473	51	196	1	5,139

Figure 2   Individual Trends in Reviews Over Time for Three Sample Books in Our Review Data Set

both their initial ratings and their long-term mean (e.g., sample book 3 in Figure 2, from roughly the first to seventh weeks). This latter observation suggests that the evidence of inefficient sales (the undershooting period) actually appears in the data (supporting H3A).

To examine the patterns found in the time series of ratings across the entire sample, we utilize an enhanced negative exponential model to fit the trend in reviews over time. The model is formulated as follows:[14]

$$AvgRating_{it} = f_0 + f_1 \exp[-f_2 \cdot T_{it}] \cos[f_3 \cdot T_{it}] + u_i + e_{it}. \quad (1)$$

$AvgRating_{it}$ represents the average rating of all reviews posted for book i between the time it was released and time t, T_{it} denotes the time difference (in weeks) between the release date and time t, u_i stands for the idiosyncratic characteristics of each individual book that are constant over time (a fixed effect), and e_{it} is a random error. Depending on the signs of

[14] We also tried other models including Box-Cox transformation and a third-order polynomial model. The enhanced negative exponential model returns a better fit.

f_1 and f_2, this model can accommodate an increasing ($f_1 f_2 < 0$), decreasing ($f_1 f_2 > 0$), or no trend ($f_1 f_2 = 0$) over time. Thus, the distinction between H1A, H1B, and H1C is simply a test of the signs of f_1 and f_2. The cosine term captures the potential for undershooting. If there is no undershooting, this model becomes a standard exponential model ($f_3 = 0$). Thus, the test of H3A is a test of the value of f_3.

We initially run the model for the entire review data set (2,203 books) using the technique described in Greene (1999, pp. 444–446) for nonlinear fixed effects models with a large cross-sectional dimension. The best-fit model (with standard errors shown in parenthetical figures under the coefficients) is estimated to be

$$AvgRating_{it}$$
$$= 3.89 + 0.42 \exp[-0.156 \cdot T_{it}] \cos[0.0001 \cdot T_{it}] + \hat{u}_i + \hat{e}_{it}. \quad (2)$$
$$\;\;(0.0016)\;\;\;(0.0041)\qquad\;\;(0.0034)\qquad\qquad(3.6830)$$

Regression results are summarized in Table 3, Column 1. This estimate suggests that the overall trends in the population are reasonably described by a standard negative exponential model. This further suggests that the predominant trend in reviews is negative, but that the undershooting period is not particularly pronounced when examined over the entire

Table 3 Regression Results for Enhanced Negative Exponential Model (1)

Covariates	Column 1 All books		Column 2 All books excluding first 5 reviews	Column 3 Books that decline w/undershooting	Column 4 Books that decline w/o undershooting
f_0/g_0	3.89***	(0.0016)	3.84*** (0.0019)	3.71*** (0.0044)	3.63*** (0.0058)
f_1/g_1	0.42***	(0.0041)	0.38*** (0.0069)	0.63*** (0.0153)	0.99*** (0.0158)
f_2/g_2	0.156***	(0.0034)	0.19*** (0.0074)	0.11*** (0.0052)	0.14*** (0.0050)
f_3/g_3	0.0001	(3.6830)	0.0001 (8.5925)	0.25*** (0.0035)	0.04*** (0.0120)
Number of obs.	82,131		77,399	19,989	45,044
Number of groups	2,203		2,203	427	1,133
R^2-within (%)	22.58		8.61	13.63	16.55

Note. Standard errors in parentheses; ***$p < 0.001$.

population. To correct for the potential forum manipulation (e.g., that the first reviews are contributed by the publisher, author, or author's friends and colleagues), we omitted the first five reviews from the data, reestimated the coefficients, and found virtually identical results (see Table 3, Column 2 for these results).

Although these population results provide support for the existence of negative trends, this aggregate analysis may mask important variations across books. Accordingly, we estimate the model (1) for each individual book in the sample. We use a grid search procedure to identify the initial parameter estimates for each of the 2,203 nonlinear fixed-effects models (see Norman and Smith 1998, pp. 516–518). These 2,203 books divide into three groups according to the presence of decreasing, increasing, or nontrending patterns to the reviews as revealed by the parameters of each individual regression. First, about 70.81% (1,560 books) show a significant (at the 10% level) decreasing trend in reviews over time. Second, about 18.20% (401 books) show a significant (at the 10% level) increasing trend in reviews over time. Third, the remaining 10.99% (242 books) show no significant pattern in reviews over time. This again suggests that most reviews show a declining trend but that all three hypothesized conditions (H1A, H1B, and H1C) appear in our data.

Ideally, we would also like to test patterns of reviews against the correlation parameter (ρ). However, because we are not able to measure correlation directly, as an alternative, for each book we proxy correlation by the number of hardback books published in the past by the author(s) of the book.[15] As discussed before, one of the reasons why early book reviews are likely to be positively biased is that the readers who tend to buy early are usually fans of the author's previous books and, thus, are more likely to enjoy the new book and write positive reviews compared to the general population. Consequently, the more books the author has published before, the more likely the early buyers of the new book are fans of the author, leading to a higher correlation between demand and quality perception. Therefore, it is reasonable to use the author's previous publications to proxy relative correlation across books and test H2. We use logistic regression to estimate the likelihood that the reviews of a book follow a declining trend over time given the number of previous books written by its author(s). Results are shown in Table 4, Column 1. For the books with productive authors (thus likely to have a higher correlation), the reviews are more likely to decline over time. This supports H2.

We also find that out of the 1,560 books whose average ratings decline over time, 27.37% (427 books)[16] show an undershooting period, with all

---

[15] The data are collected as follows: For book i, we use Amazon.com's "advanced search books" tool to search for hardback books which were published before the publication date of book i and whose author name exactly matches the name of the author (or one of the authors) of book i. We then collect the total number of books shown in the search result and repeat the process for each book in our sample.

[16] This number is a conservative estimate considering the fact that our ability to discover undershooting is limited by the statistical power of the empirical analysis (because regressions are run for each individual book instead of the entire sample).

Table 4 Regression Results for Logistic Regressions to Estimate the Probability that Reviews of a Book Follow a Declining Trend Over Time or Exhibit Undershooting
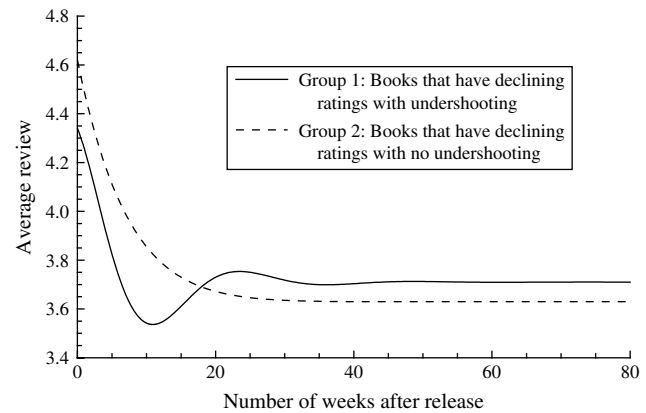
Dependent variable	Column 1 Pr{Declining trend}	Column 2 Pr{Undershooting}
Proxy for correlation (ρ)	0.005*** (0.001)	0.005*** (0.001)
Proxy for consumer heterogeneity (σ_q)	N/A	0.620*** (0.151)
Number of obs.	2,203	2,203
LR chi2	22.40***	51.14***

Note. Standard errors in parentheses; *** $p < 0.001$.

parameters individually significant at the 10% level ($f_1 > 0$, $f_2 > 0$, $f_3 > 0$). The remaining 1,113 books show declines but not significant undershooting. Estimation of the population model (1) on these two subgroups (decliners with and without undershooting) of books is summarized in Table 3, Columns 3 and 4.[17] Figure 3 shows the estimated curves traced for the coefficient values estimated. The undershooting happened, on average, 6 to 19 weeks after a book was released, and the average rating during this period is about 0.16 points lower than the long-term average. This is substantial because the average ratings across this entire population have a 0.6 standard deviation—in other words, this review bias is as large as one fourth of the standard deviation in ratings across all books we examine. As discussed in §3.2, the existence of undershooting (supporting H3A), even if for a fraction of books, rejects the hypothesis that consumers fully correct for the review bias.

To further examine whether the observation of undershooting is simply random or reasonably related to book characteristics (to test H3B), we utilize logistic regression to estimate the probability that a book exhibits undershooting controlled for both correlation and consumer heterogeneity. For each book, consumer heterogeneity is estimated using the standard deviation of all consumer ratings posted for that book, and correlation is again proxied by the number of previous books published by its author(s).

---

[17] Although f_3 is statistically significant in Column 4, its size is so small that the estimated curve (Group 2 in Figure 3) declines monotonically with no undershooting. It is possible that due to low statistical power of the individual regressions (see footnote 16), some books with undershooting are left in Group 2, which causes f_3 to be significant.

Figure 3 Estimated Trend in Reviews Over Time for Two Subpopulations

parameters individually significant at the 10% level

Results are shown in Table 4, Column 2. If undershooting is totally random, then there should be no relationship between the presence of undershooting and consumer heterogeneity or the author's previous publications. Instead, we find that books targeted to more heterogeneous consumers and written by more productive authors (i.e., likely to have a higher correlation between demand and quality perception) tend to show greater undershooting, which supports H3B.

4.3. Empirical Evidence: Impact of Consumer Reviews on Book Sales

Our previous analysis suggests that all of the hypothesized patterns described in Hypotheses H1A–H1C appear in our data, with declining reviews being the dominant trend. In addition, we found that a substantial number of books (roughly 20% of the population) show a statistically significant pattern of undershooting phenomenon (supporting H3A) that could be possibly associated with consumer welfare loss. As an alternative approach, we directly measure the potential for welfare loss by examining whether consumers respond to the time-varying (potentially biased) component of reviews rather than simply purchasing products based on overall quality (which is reflected in the long-run average). Our approach is based on prior work (e.g., Chevalier and Mayzlin 2006), except that we use slightly different control variables and estimate the model for a single retailer rather than the difference between two retailers. In addition, instead of simply incorporating ratings directly

Table 5   Description of Our Measures Used in Sales Model

Measures	Description
$SalesRank_{it}$	Sales rank on Amazon.com for book i at time t
$Sales_{it}$	Sales volume on Amazon.com for book i at time t, estimated based on the relationship of $\text{Log}[Sales_{it}] = 9.97 - 0.78\text{Log}[SalesRank_{it}]$
P_{it}	Price offered by Amazon.com for book i at time t
$NumofReview_{it}$	Number of book reviews posted on Amazon.com for book i since it was released till time t
P_{it}^c	Minimum price of book i offered by competing retailers and listed on Pricescan.com at time t
$Promotion_{it}$	$\dfrac{\text{List price} - \text{Second maximum price listed on Pricescan.com}}{\text{List price}}$ for book i at time t
$AvgRating_{it}$	Average rating of all reviews posted for book i since it was released until time t
$\bar{R}_i$	Population average rating for book i which is constant over time
R_{Tit}	Time-variant component in average rating for book i at time t
T_{it}	How long book i has been on the market since it was released until time t
$CategoryDummies_i$	Book category for book i classified by Amazon.com
$ShippingDummies_{it}$	Shipping time stated by Amazon.com for book i at time t
$WeekDummies_{it}$	The week in which the sales data were collected for book i at time t

into the model, we divide ratings into two components: a time-invariant long-term mean rating for each book and a time-varying component which captures possible self-selection bias among early buyers. To differentiate between the long-term mean and the temporal component, we utilize the estimates of model (1) on the sample population. The long-run mean component is equal to the sample mean plus the book-specific fixed effect $(3.89 + \hat{u}_i)$, denoted as $\bar{R}_i$. The bias component is equal to the time-varying component plus the random deviation from the estimated model $(0.42 \cdot e^{-0.156 T_{it}} + \hat{e}_{it})$, denoted as R_{Tit}. Thus, the sum of these two components equals the actual rating at a point in time $(AvgRating_{it} = \bar{R}_i + R_{Tit})$. Our hypotheses suggest that if consumers account for the bias in ratings when making purchase decisions, sales should be correlated only with the long-term mean (suggested by H4A), not with the temporal component (suggested by H4B).

Because we are pooling multiple books in this analysis, we need to control for heterogeneity across the sample and over time. To control for price-induced temporal variations in sales, we include four variables: the book price offered by Amazon.com (P_{it}), the best available price (P_{it}^c) from a competing seller (from Pricescan.com), and two variables that provide a measure of promotions ($Promotion_{it}$ and $Promotion_{it}^2$). Ideally, we would like our measure of promotions to control for incremental sales associated with time-limited promotions offered by publishers.

Because we cannot observe this directly, we proxy this by the difference between list price and the price of the second-highest retailer because we expect that greater numbers of retailers selling at discounts are associated with promotional activity.[18] We include this measure and its square to account for possible nonlinear relationships. We control for different demand for different types of books with a set of dummy variables for category ($CategoryDummies_i$), book availability with dummy variables for estimated shipping time as stated by Amazon.com ($ShippingDummies_{it}$), and general time-specific exogenous effects with dummy variables for calendar week because our sales data is measured weekly ($WeekDummies_{it}$). Following Chevalier and Mayzlin (2006), we also control for the number of reviews which may capture other idiosyncratic aspects of book demand not otherwise covered in our model. Finally, we include a time-trend variable (T_{it}), which captures the number of weeks since release, to ensure that we are not confounding our temporal review measure with a simple time trend. Because reviews also follow a time trend, this likely removes some of the "true" variance in our bias measure, making our results more conservative. See Table 5 for more precise definitions of the variables.

---

[18] The second-highest retailer price is used rather than the maximum price to avoid spurious values of the maximum prices reported in shopbots (some of which can exceed list price).

Overall, this yields the following estimating equation:

$$\text{Log}[Sales_{it}] = \beta_0 + \beta_{11}\bar{R}_i + \beta_{12}R_{Tit} + \gamma_1\text{Log}[P_{it}]$$
$$+ \gamma_2\text{Log}[NumofReview_{it}] + \gamma_3\text{Log}[P_{it}^c]$$
$$+ \gamma_4 Promotion_{it} + \gamma_5 Promotion_{it}^2 + \gamma_6 T_{it}$$
$$+ \gamma_7 CategoryDummies_i$$
$$+ \gamma_8 ShippingDummies_{it}$$
$$+ \gamma_9 WeekDummies_{it} + \varepsilon_{it}. \tag{3}$$

The critical test of our theory is that R_{Tit} should have no impact on consumers' purchase decisions—that is, that consumers can compensate for biased reviews to make rational purchase decisions. The test of H4A (long-term average reviews affect sales) is a test of $\beta_{11} = 0$, while the test of H4B (review bias influences sales) is equivalent to a test of $\beta_{12} = 0$.

Our model is estimated using the sales data set of 20,473 "book × week" observations for 1,290 books (see Column 1 of Table 6). Estimates for all control variables are consistent with expectations and are statistically significant—book sales decrease with the price offered on Amazon.com (P_{it}) and increase with the number of reviews ($NumofReview_{it}$), book promotions ($Promotion_{it}$), and reduced outside competition (P_{it}^C). Sales decrease over time (T_{it}). Consistent with prior work, average rating is related to sales ($\beta_{11} = 0.18$, $p < 0.01$), thus supporting H4A. More importantly, the time-varying review component R_{Tit} is also

strongly correlated with book sales ($\beta_{12} = 0.50$, $p < 0.001$), thus supporting H4B. This supports our conjecture that consumers do not fully account for review bias when making purchase decisions. For example, a positive review bias of 0.23 can increase sales by 12%. For a $17.24 book (the average price across all books in our sample), the same result has to be achieved by a $4.71 price reduction (the standard deviation of prices across all books in our sample).

To gauge the strength of our results, we also estimate the model (3) in fixed effects and weekly differences (because the frequency of observation is weekly, this is equivalent to first differencing). Both fixed effects and first differencing control for all cross-sectional variation in the data. Any of the control variables that are constant across time (including long-term average rating $\bar{R}_i$) are accommodated by the fixed effect or difference and do not appear as variables in the model. Thus, these tests are insensitive to problems of having inadequate controls for time-invariant differences across books. Overall, the temporal component of reviews is still positive and strongly significant in fixed effects ($\beta_{12} = 0.24$, $p < 0.001$; see Column 2 of Table 6). The results of the control variables are also broadly consistent with expectations, with the exception of the number of reviews. This is likely due to the nature of the fixed-effects model. The coefficient on "number of reviews" has the "wrong" sign, possibly because in the fixed-effects model the number of reviews not only captures the potential future popularity of each book but also is an indication of past accumulated sales.

There are additional considerations in interpreting the first-difference model. Because the data we use for the first-difference model are collected weekly, one potential concern is that if book sales and book ratings simultaneously decline every week, then even though the two time trends are independent, the coefficient for changes in reviews can still be significant. We rule out this possibility by incorporating the timing of the last review posted between two consecutive data collection dates into our first-difference model. We introduce an additional variable d_{it}, which represents the time distance (measured as number of days) between the date when the last review during the week was posted and the date sales rank was measured. If there are no reviews posted between any two consecutive

Table 6   **Regression Results for Sales Model**

Log[$Sales_{it}$]	Column 1 Cross-sectional sales model	Column 2 Fixed-effects sales model
Log[P_{it}]	−0.36** (0.04)	−0.38*** (0.06)
Log[$NumofReview_{it}$]	0.71*** (0.01)	−0.60*** (0.03)
Log[P_{it}^c]	0.35*** (0.02)	0.05*** (0.01)
$Promotion_{it}$	2.66*** (0.37)	−0.89*** (0.24)
$Promotion_{it}^2$	−2.05 (2.30)	4.95*** (1.49)
$\bar{R}_i$	0.18** (0.01)	N/A
R_{Tit}	0.50*** (0.04)	0.24*** (0.04)
T_{it}	−0.01*** (0.0003)	−0.03*** (0.001)
Number of obs.	20,473	20,473
R^2 (%)	30.89	18.35

Notes. Standard errors in parentheses; **$p < 0.01$, ***$p < 0.001$; coefficients for shipping dummies, category dummies, and week dummies are omitted from table.

Fridays, d_{it} is set to be zero. The rationale for this analysis is that if the time trend in sales and the time trend in reviews were independent, the intraweek timing of the last review would have no impact on sales and, accordingly, d_{it} and $d_{it} \cdot \Delta R_{Tit}$ should have no impact on $\Delta \text{Log}[Sales_{it}]$ in the first-difference model (in which Δ represents the one-week difference).

To isolate the source of variation in fitting the first-difference model, we restrict the sample to those books in our sales data set which had frequent sales rank updates (those with sales rank less than 100,000). In addition, we exclude observations for which the review is posted on the same day that sales rank data are collected because Amazon.com only reveals the date when a review is posted, not the time, and so we are not able to tell if the review is posted before the sales rank is updated. The regression results for the first-difference model are given in Table 7. Not only do changes in consumer ratings across time have a significant positive impact on changes in book sales, but the time when the review is posted between two consecutive data collection dates also matters. Both d_{it} and $d_{it} \cdot \Delta R_{Tit}$ are significant. The finding rejects the alternative possibility that the two time trends are independent, and correspondingly supports our claim that the time trend in reviews contributes to the time trend in book sales. This supplementary analysis as well as the inclusion of time trend in our levels regressions also suggest that the observed trends in reviews are not simply reflecting different arrival times of consumers with different ratings—within any given arrival point, reviews have a marginal effect

on behavior and even the timing of reviews within a week also affects sales.

These analyses collectively suggest that consumers do not fully correct the review bias in early book reviews (supporting H4B).

5. Model Extensions: Implications for Pricing, Profit, and Consumer Surplus

Our empirical results validate our assumptions underlying the theoretical model, suggesting both the presence of self-selection and a significant presence of review-bias effects (both in terms of number of products exhibiting these effects and their impact on sales). In this section, we examine further implications of our results for firm pricing strategy and the implications of review bias for profitability and consumer surplus.
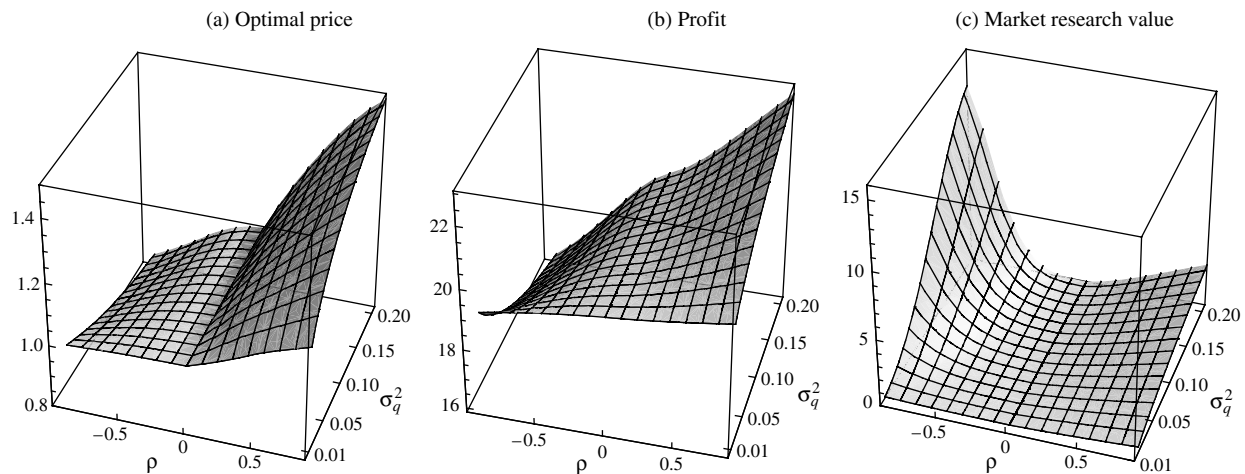
5.1. Pricing and Profit

Using the same setup described in §3, we can compute numerically how optimal price and seller profit[19] vary with different values of correlation (ρ) and consumer heterogeneity (σ_q^2). Figure 4(a) shows that if the early buyers tend to like the product, the firm has incentive to shift price upward to take advantage of the positive self-selection bias in reviews. However, if the early buyers are more likely to be critical of the product, the firm will lower price to expand the evaluation pool, exchanging losses in early periods for future profits. In addition, as shown in Figure 4(b), consumer heterogeneity increases profits only if the consumers who buy early happen to be "advocates" of the product; otherwise, consumer heterogeneity hurts because the consumers who would be willing to pay a high price under perfect information are deterred by the negative reviews. McFadden and Train (1996) suggested that learning from others hurts "niche" products. Our results further suggest that this is only true if correlation ρ is low.

Table 7   **Regression Results for Weekly Differences Model**

$\Delta \text{Log}[Sales_{it}]$	Temporal differencing model	
$\Delta AvgReview_{it}$	0.26*	(0.11)
d_{it}	−0.06*	(0.03)
$d_{it} * \Delta AvgReview_{it}$	−0.006*	(0.003)
$\Delta \text{Log}[P_{it}]$	−0.22***	(0.08)
$\Delta \text{Log}[NumofReview_{it}]$	0.0004	(0.0781)
$\Delta \text{Log}[P_{it}^c]$	−0.01	(0.01)
$\Delta Promotion_{it}$	−0.29	(0.33)
$\Delta Promotion_{it}^2$	1.50	(2.18)
Number of obs.	17,242	
R^2	0.35%	

Notes. Standard errors in parentheses; *$p < 0.1$, ***$p < 0.001$; coefficients for shipping dummies are omitted from table.

---

[19] Our model and discussion apply generally to sellers who have the ability to control price and promotional strategies. Because we do not model the interaction between publishers and resellers, we cannot make distinctions between the surpluses gained by these two groups. In general, we expect that resellers who pursue the strategies we describe will increase their profits. Whether publishers capture some of these gains depends on their wholesale pricing strategies which we do not model.

Figure 4   **Optimal Prices, Optimal Profits, and the Value of Market Research** ($K = 20$, $\tau = 2$, $q_{prior}^e = 0.3$)

(a) Optimal price   (b) Profit   (c) Market research value

These results imply that firms can encourage positive word-of-mouth in early periods by marketing the product to the proper consumer segments (those with larger ρ) or by designing the products such that the search attributes and experience attributes match the same group of consumers. That is, they should try to structure the product such that consumers who are attracted by the search attributes of the product are more likely to also appreciate the experience attributes of the product. This suggests a new perspective for firms to consider market segmentation strategy. This also reaffirms the value of honest advertising because negative reviews written by consumers who are disappointed at products that deliver less quality than the search attributes signaled can hurt firms' profits. Although these considerations were important before consumer reviews became widely available, they become more pronounced in a setting in which early reviews can have a large influence on sales.
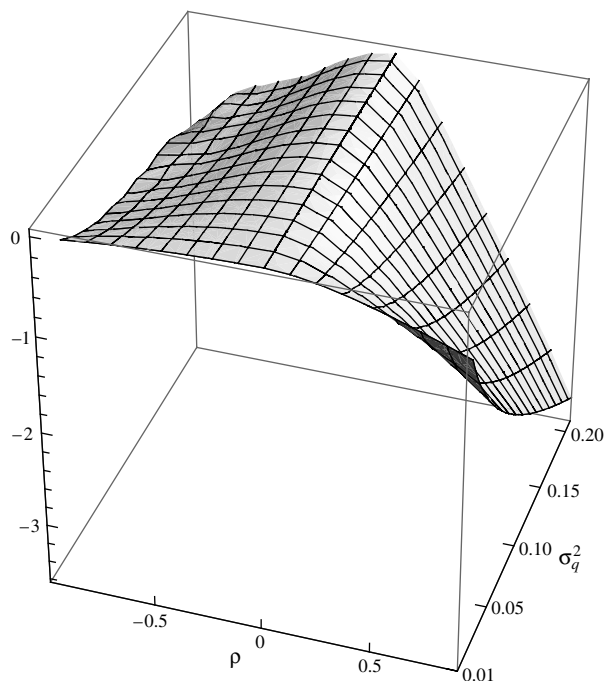
An additional strategy that firms should consider is investing in market research to better understand the likelihood of early review bias. In Figure 4(c), we compare profitability under optimal pricing, in which the firm knows the true ρ (which we have used for all results thus far), with the profitability under the assumption of $\rho = 0$ (the same "uninformedness" assumption we make for consumers). This can be interpreted as the value of conducting market research. Figure 4(c) suggests that if there is significant customer heterogeneity and negative correlation,

then market research is particularly valuable. If the firm can generally assume that $\rho > 0$, then market research is not as critical. Consequently, this suggests that it may be worthwhile, at a minimum, to engage in market research (such as product trials or focus groups) to determine if early buyers have negative bias. Firms can make adjustments to product design to reduce this bias or, if the problem is unavoidable, invite more reviewers by lowering prices to minimize the negative word-of-mouth effects. Given these alternatives, product redesign is more desirable as a remedy because products with a very negative ρ are the least profitable (see Figure 4(b)).

5.2. Consumer Surplus
If consumers are heterogeneous over their perceived quality for the same product and their tendency to buy early is correlated with their likelihood of satisfaction, then the reported reviews reflect the personal taste of the early reviewers and are consequently biased. This may mislead some consumers to purchase products they would not have bought or to pass over products they would have bought if they had corrected for reviewer bias. Figure 5 shows the utility loss that consumers incur if they follow the biased reviews, compared to the scenario in which consumers can recognize and correct the bias when they examine reviews. According to Figure 5, consumers' utility loss is more severe in the presence of positive self-selection bias. In this situation, consumers tend to

Figure 5 Consumer Utility Loss Due to Self-Selection Bias ($K = 20$, $\tau = 2$, $q_{prior}^e = 0.3$)

Table 8 Regression Results for Average Sales Model for the Books that Exhibit Undershooting

Coefficients	Average sales model
l_1	188.45*** (10.54)
l_2	0.11*** (0.007)
Number of obs.	25
R^2	97.52%

Notes. Standard errors in parentheses; *** $p < 0.001$.

pay for the overrated products, and hence are very likely to be disappointed.

Our empirical results (in §4) suggest that positive self-selection is likely present for a significant portion of books sold online (about 70%, according to our data) and that there is evidence in the form of undershooting that, for some books, this bias is large enough to cause observable purchase mistakes as signified by the undershooting period exhibited by about 20% of books. If we fit an exponential model of sales over time ($AvgSales_t = l_1 \exp[-l_2 \cdot T_t])^{20}$ for the books that exhibit undershooting (the regression results are given in Table 8) and further incorporate the trend of reviews (Column 3 of Table 3) and estimates of model (3) (Column 1 of Table 6), we can show that in the first six weeks (before the average undershooting period identified in §4.2) around 135 consumers per title who should not have bought the book made erroneous purchases because of the positive review bias. We suspect that it is these consumers who subsequently express their disappointment in reviews,

[20] Because sales at $T = \infty$ should be zero, we exclude a constant term.

which contributes to the undershooting in the review curve. If this is true, we should observe an increase in the standard variance of reviews posted shortly before the undershooting period compared to other time periods. That is, reviews contributed in this period should be different than both previous views and later purchases, increasing population variance. Indeed, we find that, on average, the variance is particularly high between the fifth and fourteenth weeks after a book is released (shortly before the average undershooting period, sixth to nineteenth weeks after release). This is consistent with the idea that a different population purchases (maybe because of the positive review bias) before the undershooting period and their disappointment brings down the average rating in the periods that follow.

This observation likely extends to other products because it is reasonable that early product adopters tend to be exceptionally favorable about a product on hard-to-observe dimensions. Because consumers' utility loss is more significant in the presence of positive self-selection bias, there could be substantial value in review systems investing to prevent or eliminate this bias. Possible solutions may include citing professional reviews, altering display order to expose consumers to varied opinions, review-weighting schemes that correct for self-selection in early periods, or encouraging reviewers to disclose more information about past purchase history or background information, thereby enabling better assessment of the reviews. How effective these solutions are in helping consumers mitigate the influence of review bias is an interesting research question for future study.

6. Conclusions and Discussion
This paper suggests that because different consumers generally hold different opinions about a product's

quality, consumer-generated product reviews may not be an unbiased indication of unobserved quality even with totally truthful reviews. If consumers' tendency to purchase early and their likelihood of satisfaction is correlated, this self-selection behavior can cause systematic bias in reviews posted in the early periods. This correlation can be positive or negative. In this study, we develop and empirically test a model that examines how idiosyncratic preferences of early buyers can affect long-term consumer purchase behavior as well as the social welfare created by review systems. The assumptions of the model are tested using online book reviews and sales data collected from Amazon.com. We find that for a majority of books, consumer reviews posted in early periods are systematically positively biased. In addition, we find evidence that consumers do not discount the early reviews when they refer to consumer book reviews for quality information.

The potential existence of self-selection bias and consumers' treatment of this bias suggest the disproportionate influence of the early buyers on market outcomes because the early reviews affect quality perception of potential buyers, and thus impact future product demand. This influence is especially pronounced for horizontally differentiated products. Accordingly, it is strategically important for sellers to manage potential self-selection bias in the market in which the consumers who tend to buy early have systematically different preferences and requirements than those of later buyers. The predictions of our analytical model suggest that the firm should adjust its product design or marketing strategies to appropriately manage potential bias reflected in early product evaluations. These strategies generally require some understanding of the behaviors of the customer population, which further suggests the value of customer relationship management systems (CRM). Not only can these systems provide the necessary signals about market demand (x_i), but they may also facilitate an even more precise strategy in which consumers with high values of perceived quality (q_i) can be targeted directly, or in which future products can be designed that will be self-selected by individuals that are both likely to have positive perceptions and likely to be influential. This is in line

with predictions about information-intensive marketing on the Internet, where "interactive online media will enable marketers to sense market forces with unprecedented accuracy and efficiency, overcoming the limitations of today's one-way research methods" (McWilliam 2000). In general, positive self-selection bias will increase seller profits, while negative self-selection bias may hurt profits but can be potentially offset by the effective use of market research and customer targeting. Thus, unless self-selection bias is completely absent, firms can benefit by adapting their strategies to take into account consumer behavior. On the other hand, the presence of self-selection bias lowers consumer surplus, which highlights the opportunity for review systems to increase consumer welfare by investing to prevent or eliminate this bias.

The model in this paper could be extended in several directions. In our current model, we are assuming that consumers' prior expectation over product quality is exogenously given. It would be interesting to take into consideration the impact of advertisements on consumers' prior expectation and examine how the firm can use the mix of advertising and pricing to balance the impacts of self-selection bias. Second, in our model, we assume that consumer reviews reflect quality rather than utility. An alternative assumption would yield interesting insights, especially in the situation in which price is not fixed over time, because reviews may be measures of utility (value of quality less price), in which case price has a direct influence on ratings. Third, the model in this paper examines the situation in which self-selection behavior in early demand drives much of the trend in reviews. If consumers' motivation to post reviews contributes significantly to the declining trend, the optimal pricing strategy may be different. How to adjust prices to influence consumers' motivation is another important issue firms should consider, especially in markets where prices change constantly. Fourth, because our model studies review bias through patterns of reviews which require a reasonable number of reviews posted over time, our results may not generalize to less-reviewed products. For those products, an interesting research question may be which consumers are more likely to be vocal and how to convert the limited number of vocal buyers to the advocates of the product. Finally, our

model does not consider the risk-reduction capabilities of online consumer reviews, which could alter demand patterns, and it does not consider optimal firm response if facing risk-averse consumers.

Acknowledgments

Appendix. Creation of the Sample

Our initial sample of books was derived from two sources:

(1) A sample of books extracted from *Books In Print* based on the following criteria:

• Publication date is between January 2000 and February 2004;

• Publication language is English;

• Book edition is hardback;

• Status is active;

• For books published before 2004, reviews are available.

(2) Hardback books that appeared at least once in the *Publisher's Weekly* bestseller lists between January 1, 2000 and February 9, 2004 and were published between January 2000 and February 2004.

Our initial sample of books was filtered to ensure a sufficient number of consumer reviews for us to study the trend in these reviews and an active sales record for us to study the impact of consumer reviews on book sales. Thus, we kept only 2,651 books in our final sample constrained by the following conditions:

For books published before 2004:

• On average, at least one review is posted on Amazon.com every 10 days, or the total number of reviews posted on Amazon.com is larger than a certain number—40 for books published in 2000, 30 for books published in 2001, 20 for books published in 2002, and 10 for books published in 2003.

• At least one review is posted on Amazon.com by the end of the first month after release.

For books published both before and after 2004:

• Sales rank is lower than 100,000.

These screening criteria are judged using the pilot data (consisting of sales and reviews data) collected from Amazon.com on February 23, 2004.

References

Akerlof, G. A. 1980. A theory of social custom, of which unemployment may be one consequence. *Quart. J. Econom.* **94**(4) 749–775.

Avery, C., P. Resnick, R. Zeckhauser. 1999. The market for evaluations. *Amer. Econom. Rev.* **89**(3) 564–584.

Banerjee, A. V. 1992. A simple model of Herd behavior. *Quart. J. Econom.* **107**(3) 797–817.

Becker, G. S. 1991. A note on restaurant pricing and other examples of social influences on price. *J. Political Econom.* **99**(5) 1109–1116.

Bikhchandani, S., D. Hirshleifer, I. Welch. 1992. A theory of fads, fashion, custom, and cultural change as informational cascade. *J. Political Econom.* **100**(5) 992–1026.

Chen, P., S. Wu. 2004. The impact of online recommendations and consumer feedback on sales. *Proc. 24th Internat. Conf. Inform. Systems*, Washington, D.C.

Chevalier, J., D. Mayzlin. 2006. The effect of word of mouth on sales: Online book reviews. *J. Marketing Res.* **43**(3) 345–354.

Clemons, E. K., G. Gao, L. M. Hitt. 2006. When online reviews meet hyperdifferentiation: A study of the craft beer industry. *J. Management Inform. Systems* **23**(2) 149–171.

Dellarocas, C. 2003. The digitization of word-of-mouth: Promise and challenges of online reputation mechanisms. *Management Sci.* **49**(10) 1407–1424.

Dellarocas, C. 2006. Strategic manipulation of Internet opinion forums: Implications for consumers and firms. *Management Sci.* **52**(10) 1577–1593.

Dellarocas, C., N. F. Awad, X. Zhang. 2004. Exploring the value of online product ratings in revenue forecasting: The case of motion pictures. *Proc. 24th Internat. Conf. Inform. Systems*, Washington, D.C.

DoubleClick. 2004. DoubleClick's touchpoints II: The changing purchase process.

Duan, W., B. Gu, A. B. Whinston. 2005. Do online reviews matter?—An empirical investigation of panel data. Working paper, The University of Texas at Austin, Austin.

Godes, D., D. Mayzlin. 2004. Using online conversations to measure word-of-mouth communication. *Marketing Sci.* **23**(4) 545–560.

Greene, W. H. 1999. *Econometric Analysis*, 4th ed. Prentice Hall, Upper Saddle River, NJ.

Kahneman, D. 2003. Maps of bounded rationality: Psychology for behavioral economics. *Amer. Econom. Rev.* **93**(5) 1449–1475.

Katz, M. L., C. Shapiro. 1985. Network externalities, competition, and compatibility. *Amer. Econom. Rev.* **75**(3) 424–440.

Lilien, G. L., P. Kotler, K. S. Moorthy. 1992. *Marketing Models.* Prentice Hall, Upper Saddle River, NJ.

Mahajan, V., E. Muller, R. A. Kerin. 1984. Introduction strategy for new products with positive and negative word-of-mouth. *Management Sci.* **30**(12) 1389–1404.

McFadden, D., K. E. Train. 1996. Consumers' evaluation of new products: Leaning from self and others. *J. Political Econom.* **104**(4) 683–703.

McWilliam, G. 2000. Building stronger brands through online communities. *Sloan Management Rev.* **41**(3) 43–54.

Norman, D. R., H. Smith. 1998. *Applied Regression Analysis*, 3rd ed. John Wiley & Sons, New York.

Piller, C. 1999. Everyone is a critic in cyberspace. *Los Angeles Times* (December 3).

Rabin, M. 1998. Psychology and economics. *J. Econom. Literature* **36**(1) 11–46.

Rogers, E. M. 1962. *Diffusion of Innovations*. Free Press, New York.

Schmalensee, R. 1982. Product differentiation advantages of pioneering brands. *Amer. Econom. Rev.* **72**(3) 349–365.

Schnapp, M., T. Allwine. 2001. Mining of book data from Amazon.com. Presentation at the *UCB/SIMS (UC Berkeley/School of Information Management and Systems) Web Mining Conf.* http://www2.sims.berkeley.edu/resources/affiliates/workshops/webmining/Slides/ORA.ppt.

Shapiro, C. 1983a. Optimal pricing of experience goods. *Bell J. Econom.* **14**(2) 497–507.

Shapiro, C. 1983b. Premiums for high quality products as returns to reputations. *Quart. J. Econom.* **98**(4) 659–680.

Simonsohn, U., D. Ariely. 2005. Non-rational herding in online auctions. Working paper, The Wharton School, University of Pennsylvania, Philadelphia.

Whitt, W. 1976. Bivariate distributions with given marginals. *Ann. Statist.* **4**(6) 1280–1289.

Villas-Boas, M. J. 2004. Consumer learning, brand loyalty, and competition. *Marketing Sci.* **23**(1) 134–145.