Marketing Science

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

Sequential and Temporal Dynamics of Online Opinion

David Godes, José C. Silva,

To cite this article:
David Godes, José C. Silva, (2012) Sequential and Temporal Dynamics of Online Opinion. Marketing Science 31(3):448-473.
https://doi.org/10.1287/mksc.1110.0653

MARKETING SCIENCE
Vol. 31, No. 3, May–June 2012, pp. 448–473
ISSN 0732-2399 (print) | ISSN 1526-548X (online)

http://dx.doi.org/10.1287/mksc.1110.0653

Sequential and Temporal Dynamics of Online Opinion

David Godes
Robert H. Smith School of Business, University of Maryland, College Park, Maryland 20742,
dgodes@rhsmith.umd.edu

José C. Silva
Fuqua School of Business, Duke University, Durham, North Carolina 27708,
josecamoessilva@alum.mit.edu

We investigate the evolution of online ratings over time and sequence. We first establish that there exist two distinct dynamic processes, one as a function of the amount of time a book has been available for review and another as a function of the sequence of reviews themselves. We find that, once we control for calendar date, the residual average temporal pattern is increasing. This is counter to existing findings that suggest that without this calendar-date control, the pattern is decreasing. With respect to sequential dynamics, we find that ratings decrease: the nth rating is, on average, lower than the $n-1$th when controlling for time, reviewer effects, and book effects. We test and find some support for existing theories for this decline based on motivation. We then offer two additional explanations for this "order effect." We find support for the idea that one's ability to assess the diagnosticity of previous reviews decreases: when previous reviewers are very different, more reviews may thus lead to more purchase errors and lower ratings.

Key words: word of mouth; online reviews; networks; Internet marketing
History: Received: August 7, 2010; accepted: March 25, 2011; Russell Winer served as the special issue editor and Bruce Hardie served as associate editor for this article. Published online in *Articles in Advance* August 25, 2011.

1. Introduction

There is a fair amount of agreement among practitioners, consumers, and the extant academic literature that, broadly speaking, *information from others matters*. Social interactions have been shown to relate to important outcomes in a range of contexts, including technology (Putsis et al. 1997), financial products (Duflo and Saez 2003), television shows (Godes and Mayzlin 2004), movies (Liu 2006, Dellarocas et al. 2007), music (Salganik et al. 2006), books (Chevalier and Mayzlin 2006), medical products (Zhang 2010, Manchanda et al. 2008, Iyengar et al. 2011), beauty products (Moe and Trusov 2011), video games (Zhu and Zhang 2010), and restaurants (Godes and Mayzlin 2009, Wojnicki and Godes 2010). Although not surprising that we would benefit from others' experiences, it is nonetheless striking the extent to which we allow ostensibly perfect strangers to influence our decisions via online reviews. A recent survey found that over two-thirds of car shoppers read online reviews before entering the dealership.[1] Approximately 68% of online shoppers read at least four reviews before making a purchase.[2] There are easily accessible online reviews for books, music, movies, television shows, consumer electronics, restaurants, hotels, airlines, cleaning services, marathons, bicycles, plumbers, doctors, summer camps, mobile calling plans, professors, and many other products and services.

Although offering the potential to assist in decision making, online reviews may in some cases present the consumer with an unwieldy glut of information. Many books on Amazon.com, for example, have hundreds, even thousands, of ratings. To make use of these opinions, given practical limits on time and cognitive processing, consumers benefit from the aggregation of the reviews into summary statistics. Review sites differ somewhat with respect to the types of summaries they make available. However, nearly all include—and feature—the mean rating. For example, Amazon.com places its "average star rating" beside the product picture on most pages. The widespread use of this data-reduction approach raises the question as to whether products can be ranked consistently according to the average rating for all, or even for average, buyers. If *Mintzberg on Management* has an average rating of 4.7 and *Peter Drucker on the Profession of Management* has an average rating of 4.3, should we conclude that the randomly drawn potential buyer should expect to like *Mintzberg* more than *Drucker*? Although the answer would be "yes" were all consumers' experiences independent, identically

[1] See McGran (2005).

[2] See Kee (2008).

drawn (iid) draws, this may not be the case. In particular, there are potentially many sources of nonindependence along the sequence of reviews. As a result, one may need to condition on these dynamic factors to consistently predict utility. In this paper, we investigate (i) whether such dynamic effects exist, (ii) on what dimensions we may expect to see dynamics occur, and (iii) several possible explanations for these effects.

We begin our analysis by demonstrating that there are actually multiple and distinct underlying dynamic processes. In particular, we show that there exist both sequential and temporal processes: *ratings change systematically over both order and time*. This is noteworthy because the existing studies have conceptualized dynamics as occurring solely as a function of one factor, although this factor has varied. Li and Hitt (2008), for example, use only a temporal explanatory variable, whereas Wu and Huberman (2008) use only a sequential one. However, neither controls for the other. This has important implications for our understanding of the phenomenon. Fundamentally speaking, a sequential process implies that the arrival of the reviews themselves drives the decline. On the other hand, a temporal process (controlling for sequence) would suggest that the reviews themselves are not the driving force that must exist independently. Because the existing literature has focused on just one process to the exclusion of the other—Li and Hitt (2008) on time, Wu and Huberman (2008) on order—there exists the very real possibility that one of them is picking up spurious effects given the potential for correlation between time and sequence. Our analysis—which shows that this is not the case—thus demonstrates that any future investigation into the dynamics of reviews must account explicitly for *both* processes.

With this new insight—and with the caveat that we must simultaneously consider both processes—we investigate more deeply each of these underlying processes: Why do they occur? We focus first on temporal dynamics. We find that even conditional on order, ratings seem to decline over time. This is intriguing because one would expect that the arrival of information—off-line word of mouth, critics' reviews, etc.—ought to result in better choices and, thus, we would expect, higher ratings. Li and Hitt (2008) argue that this decline is driven by the fact that, on average, the more people value a product, the earlier they purchase it. We offer an alternative explanation that is somewhat simpler: there is a broader underlying temporal process at work in the data. That is, over the roughly 10–15 years of online reviews, reviewers seem to have become more critical and more negative. Whereas Li and Hitt's (2008) theory is a book-level theory—there is a new "ordering" for each book—this alternative theory is a macro theory, affecting the ratings given to all books (and possibly other products). We demonstrate this by including a flexible time trend in our basic specification and find that the average ratings for all books seem to be declining over time, particularly over the most recent three-year period in our data set (2003–2005). Moreover, once we control for this macro trend, the residual temporal process is in fact *positive*: conditional on the year a review was written, ratings at the book level increase over time, as one may have expected ex ante.

We then turn our attention to the sequential process and begin with the motivation-based theory of Wu and Huberman (2008): because reviewing is costly, people are less motivated to post additional positive reviews for already highly rated books. We note that although highly compelling and logically intuitive, the published work offers little in the way of a rigorous test. Thus, we subject the theory to a direct test of one of its key implications. We are able to find evidence for the idea that the declining trend should be most pronounced for "high-cost" reviews. We then offer and analyze a complementary explanation for the sequential process: purchase errors may increase as more reviews arrive. These errors, in turn, lead to lower ratings. We hypothesize and test two possible sources for this increase in purchase errors: (a) increasing difficulty in diagnosticity assessment and (b) decreasing similarity among reviewers over the sequence. To test these, we construct a measure of similarity among reviewers. We find support for the former process by demonstrating that the sequential decline in ratings is most pronounced when reviewers are highly heterogeneous. When a reviewer is similar to those that preceded her in the sequence—implying that their reviews are "representative" of her tastes—the decline is mitigated.

The rest of this paper is organized as follows. Following our review of the literature on online reviews in §2, we describe the data we use for our analysis in §3. In §4, we present our initial evidence that there exist both sequential and temporal dynamics. Our investigations into temporal dynamics and sequential dynamics are presented in §§4.2 and 5, respectively. Finally, we conclude in §6 and discuss the limitations of our study, as well as the rich opportunities our results offer for future research.

2. Related Literature

Given the broad impact and relatively recent advent of online ratings, it is not surprising that a wide variety of disciplines—including, for example, marketing, computer science, information systems, economics, and physics—have addressed various aspects of the phenomenon. We review the literature in two categories: (i) the impact of online ratings and word of mouth

(WOM) generally and (ii) the antecedents of online ratings.

2.1. The Impact of Online Reviews

Online ratings have consistently been shown to have an impact on sales, although there exists some disagreement surrounding the question of what dimension or metric associated with the ratings is most impactful. A number of papers (Chevalier and Mayzlin 2006, Dellarocas et al. 2007, Chintagunta et al. 2010) find that positively valenced reviews increase sales. Liu (2006), on the other hand, finds that valence does not matter for movie sales but that the volume of online ratings does. Duan et al. (2008a) yield a similar finding and argue that this is evidence that online ratings are not *persuasive* (because ratings valence does not seem to have an impact) but serve instead to increase *awareness* of movies. Moe and Trusov (2011) and Zhu and Zhang (2010) find evidence for the role of both valence and volume of online reviews in models of sales. Kumar and Benbasat (2006) create a lab experiment using real-time–filtered Amazon.com data and demonstrate that the mere act of allowing for recommendations and reviews improves the perceived usefulness of the site. Thus, they demonstrate the impact not of online ratings directly but of the firm's decision to enable the provision of these ratings. Finally, Forman et al. (2008) find that a reviewer's decision to reveal his or her identity information (his or her real name, his or her location) may moderate the reviewer's impact: when this information is revealed, the impact of a review on sales is higher.

While studying the impact of online reviews on sales—or more broadly speaking, on consumer choices—a number of researchers have allowed for the possibility that this impact may, in fact, vary dynamically. Of these, most studies have reported that the impact seems to wane over time (Godes and Mayzlin 2004, Moul 2007, Hu et al. 2008, Duan et al. 2008b). One noteworthy exception in this regard is Zhu and Zhang (2010), who find that online reviews for video games are less influential in the early stages of a product's life. They conjecture that this result may be because firms in the entertainment industry promote heavily at product launch. As a result, consumers may depend less on reviews until the impact of these promotions dissipates.

The literature also contains a number of studies that have investigated the dynamic impact of WOM more broadly. On one hand, we have the important research stream on herds and cascades (Banerjee 1992, Bikhchandani et al. 1992). According to these models—and to the empirical and experimental literature that has followed (for example, Çelen and Kariv 2004, 2005; Duan et al. 2009)—individuals may optimally ignore their private information and instead follow the observable choices made by those before them in a sequence. Although these models, by design, have no word-of-mouth communication per se, the dynamic pattern of the observational learning is noteworthy in that it may be interpreted to imply that its impact is invariant over time.[3]

On a related note, there is an emerging literature in which the relative impact of WOM on one hand and the relative impact of traditional marketing vehicles on the other are assessed. Trusov et al. (2009) study the evolving membership in a social networking site over three years. They estimate the elasticity of membership growth with respect to both traditional marketing and WOM referrals, and they find that the impact of the latter remains in force for approximately three weeks. The impact of traditional marketing, on the other hand, disappears within a few days. Villanueva et al. (2009) study a Web hosting company and compare the value of customers attracted via traditional marketing methods with those attracted via WOM referrals. They find that there are important dynamic differences in the impact these marketing vehicles have on customer lifetime value. Most notably, they find that the customers acquired via traditional marketing contribute more to the firm's profits in the short term but that the opposite is true in the long term, where customers acquired via WOM contribute more. These studies may be interpreted as being consistent with the findings of Zhu and Zhang (2010) in that the *relative* impact of WOM on profits may grow over time.

2.2. Drivers of Online Reviews

Whereas there is a fair amount of agreement regarding the impact of WOM and online ratings, the literature is far less clear with respect to what drives these ratings themselves. Liu (2006) shows that the primary antecedent of WOM following the release of a movie is the volume of the previous week's WOM. The valence of the previous WOM, critics' opinions, and the existence of a star in the movie's cast did not have a significant impact. Dellarocas and Narayanan (2006) show that "ratings density"—the proportion of people who purchased the product that have rated the product online—is increasing in both the firm's marketing expenditure and disagreement among critics. They also find, consistent with Anderson (1998), that this density metric is U-shaped in quality. Finally, Qu et al. (2008) investigate the driving factors behind the ratings of Yahoo! merchants. They find that these ratings are increasing in the merchant's post-transaction

[3] Specifically, previous choices may land the entire sequence that follows in a "cascade," which is effectively impossible to exit, barring additional information. Thus, once in a cascade, the nth player's observation of player $(n-1)$ has precisely the same impact on her decisions as does the $(n+t)$th player's observation of player $(n+t-1)$.

behavior (for example, post-transaction spam, on-time delivery, and effectiveness of customer service).

Another important factor in the ratings creation process is the purchase process itself. It is notable that a number of researchers have studied the potential for self-selection in the purchase decision, including the timing of purchase and, thus, review (Li and Hitt 2008, Moe and Schweidel 2012, Hu et al. 2009, Moon et al. 2010). Indeed, this phenomenon has been suggested as a possible explanation for the common existence of a "J-shaped" distribution of posted ratings (Moe and Schweidel 2012, Hu et al. 2009).[4] On a related note, the variance of the ratings distribution is a focus of Sun (2010), who argues that demand is a U-shaped function of the variance of the ratings distribution. In particular, when the average rating is low, high ratings variance may be beneficial to the product.

Conditional on purchase, one then needs to decide whether or not to post a review. Wu and Huberman (2008) make the argument that reviewers are motivated by the expected impact their review will have on the average rating and, implicitly, on the actions or preferences of others. They hypothesize that buyers are most likely to post their reviews when the expected impact is high: when there are fewer existing reviews and/or their experience deviated greatly from the prevailing average rating. Moe and Schweidel (2012) present a joint model of a reviewer's decision about (i) whether to review and (ii) what rating to assign, conditional on a review. Moreover, their model allows for interpersonal differences in these behaviors. They find significant heterogeneity with respect to consumers' desire to post in high-consensus versus high-variance environments.

There is also growing research interest around *what* is posted and how that process is a dynamic one. Indeed, as shown by Moe and Trusov (2011), the apparent existence of a dynamic ratings generation process implies that ratings have not only a direct effect on product sales (Chevalier and Mayzlin 2006) but also a significant indirect effect via future ratings as well. Of specific interest here is the growing evidence that online ratings are characterized by, on average, a declining dynamic trend (Li and Hitt 2008, Wu and Huberman 2008, Moe and Trusov 2011, Moon et al. 2010). There are two prevailing explanations that have been offered for this phenomenon, one based on the ratings environment and the other based on self-selection independent of the ratings environment. With respect to the former, Wu and Huberman (2008) present a theory based on the decision by a potential reviewer about when, and when not, to post one's opinion. They argue that, before adding a review, one considers whether the *impact* of the review will outweigh the *cost* of submitting it. A review's impact is increasing in the deviation of one's opinion from the prevailing average and decreasing in the number of reviews already posted. When the prevailing average is high—as is often the case empirically—this would imply that observed ratings will, on average, decline. Simply put, adding another 5-star rating to a book with a thousand reviews and an average of 4.9 is unlikely to have an impact. However, adding a 1-star rating to the same book would have a larger impact and may thus be worth the "cost" of submitting it. Similarly, recent work by Moon et al. (2010) conjectures that endogenous arrival explains the negative coefficient on their dynamic covariate.

On the other hand, Li and Hitt (2008) argue, and present evidence for the idea, that the decline is caused by *self-selection* at the purchase stage: customers who place a higher valuation on a product purchase it earlier. Loyal fans of Mintzberg, for example, buy his books first. Casual fans buy next, and those who are more dispassionate buy later. This explanation for the declining average trend would seem to be based on assumptions that (a) expected utility is, on average, a good predictor of experienced utility, at least in an ordinal sense; and (b) ex post ratings reflect actual experienced utilities. Using book ranks as proxies for sales, they also conclude that consumers do not take this dynamic pattern into consideration when making purchase choices.

There are a number of critical distinctions between these two theories themselves, on one hand, and between both theories and ours, on the other. First and foremost, we focus on the underlying dynamic variable. Self-selection, according to Li and Hitt (2008), occurs over *time* and is not due directly to the arrival of reviews. Accordingly, the authors estimate a model with a time trend but no ordinal effect. Wu and Huberman (2008), on the other hand, see the previous reviews as the direct cause of the declining trend, and thus they specify a model in which the core dynamic variable is an *ordinal* one: At what location in the sequence of reviews did one arrive? We address here the relative veracity of these two proposed mechanisms—time and order—and the mechanisms the models have been specified to capture. Does the sequence of online ratings decline over time, order, or both? It is critical to note that this is hardly just an econometric exercise; in fact, it bears directly on the underlying theories. This is the first study, to our knowledge, that explicitly accounts for—indeed, demonstrates the existence of—both dynamic effects. It is clear, given our results, that future research on the dynamics of online reviews must account for both.

Second, given the existence of both effects, we delve more deeply into the explanations for both. We contribute to the literature on temporal effects by investigating a potential alternative explanation to that of Li

[4] In this regard, it is worth noting that Lorenz (2009) finds that the assumption of U-shaped histograms is not universally true.

and Hitt (2008): a macro-level negative trend. That is, we show that outside of the reviews for any given book, the average reviewer is becoming more critical. Indeed, after controlling for this process, the temporal trend observed by Li and Hitt (2008) is reversed. Finally, we investigate more deeply the ordinal process. We first provide an explicit test of the theory of Wu and Huberman (2008). To our knowledge, this is the first such test. We find evidence for their theory in that the sequential decline seems most pronounced when the cost of posting a review is highest. We then propose an alternative theory based on the idea that reviews, in the aggregate, may become less useful as more arrive. To test this, we construct a measure of similarity among reviewers and find that—after controlling for individual positivity and the costliness of the review (Wu and Huberman 2008)—the declining trend exists only when reviewers are highly dissimilar from those that preceded them. Indeed, the trend is reversed when this dissimilarity is sufficiently low.

To summarize, we see the contribution of this paper relative to the existing literature to be five-old: (1) We demonstrate that there exist multiple and distinct dynamic processes affecting simultaneously the ratings distribution. These processes should all be accounted for by future researchers. (2) One of these processes is a sequential one that seems to be driven in part by the degree of similarity among reviewers. When one is very different from previous reviewers, their reviews are more difficult to evaluate and thus less useful, leading to more purchase errors. This theory has not been raised or tested previously in the literature. (3) We provide the first rigorous test of the motivation-based theory offered by Wu and Huberman (2008) and find support for their theory. (4) We find that the temporal trend studied by Li and Hitt (2008) is more complicated than it appears. Specifically, we find a general negative trend independent of any book but a positive trend along each book's life. This would seem to be inconsistent with their theory but consistent with the idea that the arrival of exogenous information should result in better decisions and higher ratings. (5) Finally, we show that the similarity of adopters declines over the sequence. Although a peripheral finding, we see it as a novel one worthy of future inquiry. Please see Appendix A for a summary of the most closely related existing work.

3. Data

Our primary empirical analysis utilizes book review data gathered from Amazon.com. We collected all available reviews for each of the top 350 selling titles as of noon, October 27, 2005. For each review, we collected the date, the reviewer's name, the length of the review text, and the star rating (an integer between 1 and 5). Our initial data set included 76,657 review observations. We immediately removed 1,907 observations that were duplicates in the data set because Amazon.com allows for multiple versions of the same book to be listed separately although their set of reviews is identical. Using these data, we form for each review the variable *STARS*, the number of stars assigned by the reviewer, and *TIME*, the number of days since the first review was posted for the book. This latter measure represents our primary temporal variable.

A critical distinction between our data set compared with those used by previous researchers (Chevalier and Mayzlin 2006, Li and Hitt 2008, Wu and Huberman 2008) is that to (a) control for reviewer-level characteristics and (b) assess the (dis)similarity among reviewers, we also collected, for each reviewer who wrote at least one of the reviews in the main data set, all of the *other* reviews they had written as of October 27, 2005. To compile this reviewer-level data set, we collected an additional 514,381 reviews. To reiterate, while other researchers have collected "longer" data sets, with more book-level observations, our efforts to collect more data at the reviewer level allow us to control for observed reviewer heterogeneity, which other researchers have not been able to do. As will be shown, this contributes a great deal of information to our models.

To assess the impact of sequential dynamics, we form the variable *ORDER*, which captures the position of a review in the sequence of reviews for a given book. One challenge we face in this regard is that the timing of each review is available only at the date level; at the time of our data collection, Amazon.com provided no publicly available time stamp. Thus, we cannot identify the order of reviews that arrived on the same day. To deal with this, we allow for "ties" in the sequence value such that multiple reviews for a given book may have the same ordinal sequence value. Thus, for example, there may be several "547th" reviews for a certain book even though, in reality, they arrived at a different time on the same day. We account for these ties in our definition of the *ORDER* variable as follows. Let d_r capture the day on which review r arrives. Again, there will be multiple reviews with the same value of d for a given book. For each d' in the data set, we then form $\Delta_{d'} \equiv \{r\colon d_r = d'\}$, the set of reviews for which $d_r = d'$. We then form, for each d', the variable *ORDER* as follows:

$$ORDER(d') \equiv \sum_{d<d'} N(\Delta_d) + 1,$$

where $N(S)$ is the cardinality of set S. Therefore, practically speaking, this method assigns the same sequential position to all reviews that arrived on the same day. Moreover, it preserves ordering in the face of ties in that the sequential position assigned to a review is always equal to 1 plus the number

of reviews that arrived on all previous days. As an example, imagine that one review arrived on January 1, two arrived on January 4, and one more arrived on January 10. According to this approach, the first review is assigned $ORDER = 1$, the second two receive $ORDER = 2$, and the final review receives $ORDER = 4$.[5] Econometrically speaking, this means that we do not have a true dynamic panel data set because we have multiple observations for many sequential positions.

Reviewers are likely to differ in their approach to assigning ratings to books: some may be more "positive" than others. We control for this by forming a reviewer-level average rating. Stated formally, let i be the author of review r for book b. Then the variable $REVAVG_{irb}$ for this review is the average of all of reviewer i's reviews on books other than b. Note again that most of these books are not in the data set of 350 books on which our main estimation takes place. Note also that were reviewer i to write reviews for books b and b' in our data set, the two reviews may be assigned different *REVAVG* variables because they would be based on a different set of books; the former would include b' but not b, whereas the latter would include b but not b'. We view the data collection in support of the *REVAVG* variable as crucial to our analysis because it accounts for what one may expect to be a large amount of individual-level heterogeneity. This is particularly important here, given that reviewer-level fixed or random effects are not an option for us because many reviewers authored only a single review in our core data set of 76,657 reviews.[6]

Although our use of this reviewer-level covariate is straightforward and assists in controlling for an important source of variation, it also raises a challenge. To create the measure, we rely on the fact that many reviewers are willing to associate a consistent identifier with each of their reviews. This identifier may be a real name—which Amazon.com verifies using a credit card—or a pseudonym. However, many reviewers are unwilling to provide either. When this is the case, we cannot form a reviewer average because we cannot link a reviewer's review to others outside of our data set. A similar problem occurs when a reviewer with an identifier has reviewed only one book. Again, this precludes our formation of a reviewer-level average. Our solution to this in the core analysis is to drop these reviews from the data set.

[5] Because we will investigate the extent to which one is affected by previous reviews, this approach would seem to be superior to alternative approaches such as, for example, assigning the average $ORDER = 2.5$ to the two January 4 reviews.

[6] Specifically, fewer than 6,500 of the reviewers wrote more than one review. Fewer than 2,500 wrote more than two. Recall, however, that our *REVAVG* measure identifies other reviews that these reviewers wrote for books outside of our data set.

Table 1 **The Impact of Deleted Reviews**

Star rating	Used in analysis	Full	Chevalier and Mayzlin (2006)
1 stars (%)	8	11	9
2 stars (%)	6	6	8
3 stars (%)	10	8	11
4 stars (%)	20	16	20
5 stars (%)	56	59	53
Avg stars	4.09 (1.27)	4.08 (1.37)	4.14 (0.70)

Note. Values in parentheses are standard deviations.

This allows us to control for individual-level reviewer characteristics via *REVAVG*. The number of reviews for which these issues are present is not trivial. After discarding these two sets of reviews, we are left with 34,747 usable reviews. As above, we preserve the ordinality of the review sequence following the discarding of reviews. For example, if the 374th review for *Toxic Bachelors* has an anonymous reviewer, then it is discarded. However, the *ORDER* variable for the previous and following reviews are still assigned values of 373 and 375, respectively.

Given how many reviews are being discarded, it is important to assess the extent to which the data we use in the analysis are drawn from a similar distribution as those discarded. The mean of *STARS* for the full data set is 4.08 (SD = 1.37), whereas that for the retained set of 34,747 is 4.09 (SD = 1.27). Table 1 provides a comparison of our data set, containing both the initial 74,750 and the 34,747 reviews we employ for the main analysis, with that used in Chevalier and Mayzlin (2006). The table suggests that the retained reviews have a higher central tendency than the full data set. Nonetheless, it appears that at the level of quantitative rating, the distribution with which we are working is quite similar to that investigated by Chevalier and Mayzlin (2006), who utilized a random sample of books. Although this should provide some comfort that our decision to discard more than half the data set does not bias our results, we demonstrate more directly in §4 that the discarding of these reviews does not have an impact on either the qualitative nature or the strength of the core dynamic phenomena (see Models (2) and (3) in Table 4).

As a final note on the limitations of our data set, we do not have access to sales data. Previous researchers (Li and Hitt 2008, Chevalier and Mayzlin 2006,

Table 2 **Summary Statistics**

Variable	Mean	SD	Min	Max
STARS	4.09	1.27	1	5
REVAVG	4.11	0.88	1	5
ORDER	570.80	772.33	1	4,982
TIME	877.47	915.88	1	3,645

Table 3 Pairwise Correlations

	STARS	*REVAVG*	*TIME*
REVAVG	0.29		
TIME	0.04	0.03	
ORDER	0.01	0.00	0.10

Chevalier and Goolsbee 2003) have made use of a log-transformation of the sales rank as a proxy for sales volume. However, given that our analysis is retrospective and that the time series of historical ranks are not generally available, such an approach is not practical in our case. Thus, we control for vertical book quality using book-level fixed effects.

Table 2 presents the summary statistics of the main variables of interest. Table 3 contains the pairwise correlations.

4. Characterizing the Dynamic Process for Online Ratings

We begin our analysis by investigating the existence and nature of the dynamic process or processes at work in online ratings data. A striking distinction between the models offered by Li and Hitt (2008) and Wu and Huberman (2008) is that the former uses *time* as its primary dynamic variable, whereas the latter looks at the *sequential position* of the review—how many reviews have already been submitted at time t. Importantly, neither of these studies conditions on the other dynamic variable. Thus, it is not clear that they are studying different dynamic processes at all or, perhaps, different proxies for the same process. As a first step in our analysis, then, we are interested in establishing the extent to which a dynamic pattern exists and—if so—whether it is primarily a temporal process and/or a sequential one. We argue that the distinction between time and order is not just intellectually interesting—it could have important practical implications as well. A primarily sequential process would suggest that the number of previous reviews has an impact on an individual's rating. This would be consistent with the theory of Li and Hitt (2008) but not that of Wu and Huberman (2008). On the other hand, a temporal dynamic process would imply that the ratings generation process is driven by something outside of the review environment. That is, if after controlling for the number of reviews that have arrived, the conditional distribution of ratings changes systematically as a function of time, it would suggest that there is some aspect of the passage of time other than the arrival of more reviews that is driving the dynamic effect. This would be consistent with the Wu and Huberman (2008) theory but not with Li and Hitt (2008).

As a first look at the data, we present in Figure 1 the data aggregated across reviews for each value of *ORDER* and *TIME*. Although not a substitute for formal models, which we present below, these panels are strong indications that there are dynamic effects at work in the ratings data. Moreover, it is important to note that Figure 1 also demonstrates that the effects we study here are nonnegligible, having a substantive impact on both the conditional average of new ratings (panels (a) and (c)) and the conditional cumulative average (panels (b) and (d)).

Stated more formally, we model the ratings generation process at the review level as a function of how much time has elapsed since the first review (*TIME*) and the position of the review in the sequence of reviews for the book (*ORDER*). Distinct from the existing literature (Li and Hitt 2008, Wu and Huberman 2008), in our main models we also control for both observed reviewer-level and unobserved book-level heterogeneity via *REVAVG* and book fixed effects, respectively. Our primary dependent measure is $STARS_{ib} \in \{1, 2, 3, 4, 5\}$, the rating assigned by reviewer i to book b. Five stars is the best rating; one star is the worst. Given the discrete and ordered nature of this variable, the appropriate model is the ordered-logit model. The following is our focal specification for the reviewer's latent evaluation U_{ib}:

$$U_{ib} = \beta_1 \cdot TIME_{ib} + \beta_2 \cdot ORDER_{ib} + \beta_3 \cdot REVAVG_{ib} + \sum_{b=1,\ldots,349} \delta_b + \varepsilon_{ib}, \qquad (1)$$

where δ_b is the fixed effect for book b. This equation captures the idea that reviewer i's latent evaluation for book b can be explained by the book's vertical quality δ_b, by the reviewer's idiosyncratic approach to rating as captured by the average of her other reviews, and by where she arrives in the purchase diffusion in both time and sequence. The model is estimated via maximum likelihood as part of which—in addition to the main parameters of interest (the βs and δs)—a set of four cutoff values μ_k, $k \in \{1, 2, 3, 4\}$ are estimated such that the discrete ordinal response $STARS_{ib}$ is generated based on where the latent evaluation U_{ib} falls vis-à-vis the cutoffs:

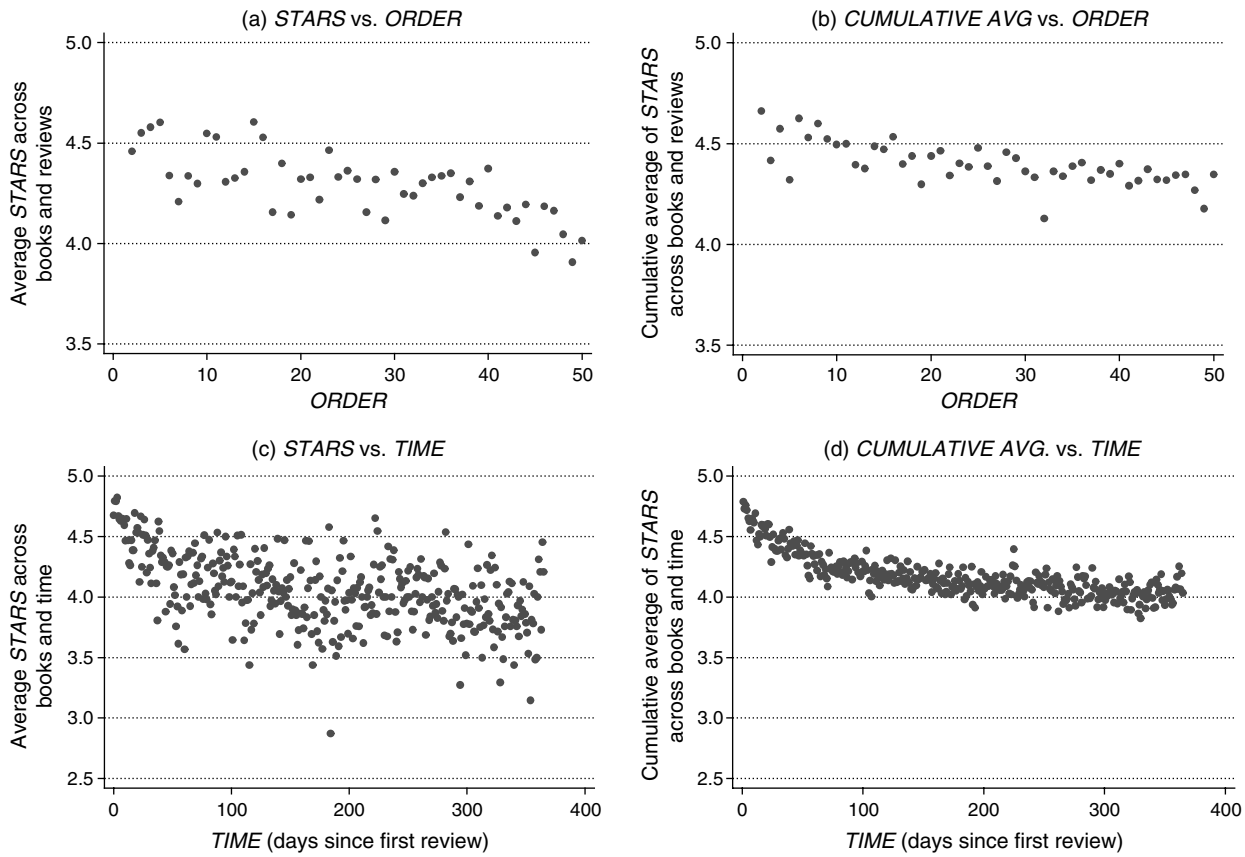
$$\begin{aligned} STARS_{ib} = 1 \;&\Leftrightarrow\; U_{ib} < \mu_1, \\ STARS_{ib} = k \in \{2, 3, 4\} \;&\Leftrightarrow\; U_{ib} \in [\mu_{k-1}, \mu_k), \\ STARS_{ib} = 5 \;&\Leftrightarrow\; U_{ib} \geq \mu_4. \end{aligned} \qquad (2)$$

Table 4[7] presents our estimation results, in which we assess the existence of these two distinct dynamic processes.

All models control for book effects, whereas Models (1) and (4) control for reviewer effects as well. Our

[7] The estimates of the cut points μ_k are suppressed for reasons of exposition. Note that, in all the tables presented in this paper, all cut points are significantly different from the adjacent cut point(s) at the $p < 0.001$ level.

Figure 1 **A "Model-Free" View of the Data**



Notes. Panels (a) and (b) depict the evolution of ratings over the sequence of reviews. Each point in panel (a) represents the average *STARS* rating across all books and reviews at the value of *ORDER* on the *x* axis. Thus, for example, the first point is the average *STARS* rating for all reviews that are the first in the sequence. To partially control for quality effects, we restricted the sample to those books with at least 50 reviews such that all points are averages over the same set of books. Panel (b) uses the same sample and presents the *cumulative STARS* average. Panels (c) and (d), on the other hand, present the average *STARS* rating and cumulative average *STARS* rating, respectively, for all books that had reviews at least one year after the first review. Again, we do this to ensure that each point represents an average over the same set of books.

main result is in Model (1). Here, we establish that there seem to be two separate and distinct dynamic processes at work: ratings appear to decline in both *TIME* and *ORDER*, each conditional on the other. Thus, holding constant how long it has been since the book's first review—and, thus, by implication, systematic changes in expectations by prospective buyers conditional on the observable attributes—ratings are lower the more reviews that have already been provided. Analogously, holding constant the number of reviews—and, thus, the average systematic impact additional reviews have on one's ability to choose products that deliver high utility—ratings decline in the amount of elapsed time since the book has been reviewed.

The coefficient estimates for *REVAVG* suggest that a reviewer's individual rating tendency has a great deal of explanatory power in this model: some reviewers assign systematically higher ratings to their purchases than do others. Although not a focal element of our research, we nonetheless see the identification of the source(s) of this individual effect as an interesting question for future analysis. For example, one might imagine at least two possible explanations. On one hand, it may be because some consumers are "easier graders" than others. On the other hand, it may also be the case that some consumers are better at predicting their own utility and, thus, purchasing products that more closely match their preferences. The relative prevalence of these two sources of heterogeneity—scale usage and purchasing ability—would likely have important theoretical and managerial implications.

The results in Model (1) are consistent on the surface with those reported by Li and Hitt (2008) in that ratings appear to decline in time on average. Although Li and Hitt (2008) control for quality, they do not control for reviewer characteristics or sequential position, and these results thus serve as a robustness check by ruling out potential alternative explanations. Model (1) suggests that, for example, the time-based ratings decline is not due to easier graders entering the market earlier

Table 4 Estimation Results

	Ordered logistic regression of *STARS*					
	(1) Temporal and sequential effects	(2) Full sample	(3) Chevalier and Mayzlin (2006) random sample	(4) Review-year effects	(5) Review-year effects in full data set	(6) Review-year effects in Chevalier and Mayzlin (2006) data set
REVAVG	6.56E–01***			6.56E–01***		
	(1.49E–02)			(1.49E–02)		
TIME	−1.22E–04***	−2.43E–04***	2.50E–04***	2.53E–04*	2.54E–04**	2.80E–04***
	(3.10E–05)	(1.75E–05)	(6.24E–06)	(1.21E–04)	(8.20E–05)	6.44E–06
ORDER	−2.03E–04***	−7.85E–05***	−1.15E–03***	−2.07E–04***	−9.49E–05***	−8.65E–04***
	(3.07E–05)	(1.78E–05)	(3.30E–05)	(3.20E–05)	1.86E–05	3.50E–05
1995					1.80E+01***	2.65E–01
					(1.01E+00)	(3.11E–01)
1996					8.59E–01*	2.74E–01***
					(3.65E–01)	(5.80E–02)
1998				−1.33E–01	2.85E–02	−1.03E–01***
				(4.14E–01)	1.21E–01	(2.52E–02)
1999				−3.31E–01	−4.25E–01***	−2.51E–01***
				(3.95E–01)	(1.22E–01)	(2.36E–02)
2000				−5.69E–01	−6.92E–01***	−2.78E–01***
				(3.90E–01)	(1.37E–01)	(2.35E–02)
2001				−8.21E–01*	−9.55E–01***	−3.30E–01***
				(4.04E–01)	(1.57E–01)	(2.33E–02)
2002				−8.39E–01*	−1.09E+00***	3.94E–01***
				(4.23E–01)	(1.79E–01)	(2.32E–02)
2003				−1.04E+00*	−1.25E+00***	−4.33E–01***
				(4.48E–01)	(2.06E–01)	(2.46E–02)
2004				−1.17E+00**	−1.37E+00***	−4.33E–01***
				(4.70E–01)	(2.28E–01)	(2.94E–02)
2005				−1.26E+00**	−1.49E+00***	
				(4.95E–01)	(2.53E–01)	
Book effects	Fixed	Fixed	Random	Fixed	Fixed	Random
N	34,747	74,750	256,911	34,747	74,750	256,911
LL	39,305	85,016	1,397,858	39,296	84,964	1,400,709
AIC	79,247	170,677	2,795,730	79,246	170,593	2,801,449

Notes. Robust standard errors are in parentheses. LL, log likelihood; AIC, Akaike information criterion.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

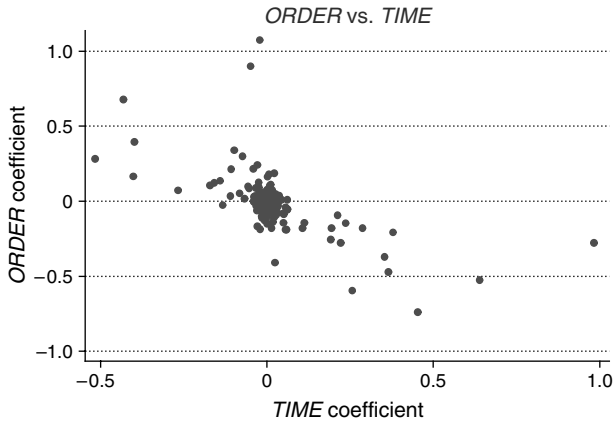
or to a sequential (as opposed to time) effect. Analogously, Model (1) serves as a robustness check on the Wu and Huberman (2008) results by controlling for quality, reviewer effects, and time, for which their univariate analysis does not allow. Their results persist here: on average, book reviews decline over the sequence of reviews even after controlling for all of these factors.

One interpretation of the results in Model (1) is that each person is affected by both temporal and sequential factors. Because these effects are averaged over both books and individuals, however, it is also possible that these two processes reflect the impact of dynamics on distinct consumer segments. Van Den Bulte and Joshi (2007), for example, specify a model with "influencers" and "imitators." In their model, influencers make their own decisions based on expected-utility maximization, whereas imitators are influenced by them. For example, they may read the influencers' reviews online and then decide whether or not to buy. This model would be consistent with our results in that the influencers may be making decisions that exhibit temporal dynamics, whereas the imitators' decisions reflect primarily sequential dynamics because they are affected by the previous choices of others. We see the investigation into this question as a potentially fruitful area for future research.[8] See, for example, Moe and Schweidel (2012) for an analysis of latent segmentation along these lines.

It is important to note that our analysis in Table 4 provides estimates of the dynamic effects that are averaged across both reviewers and books. It would be unwarranted to conclude from these results that ratings for *all* books will decline over *TIME* and/or *ORDER*. The results show that this will occur only on average. Although it is difficult to estimate within the classical framework we employ here the extent

[8] For another example of the simultaneous estimation of both time and order effects, see Karshenas and Stoneman (1993).

Figure 2 **Scatterplot of Coefficients on *TIME* and *ORDER* on the *x* and *y* Axes, Respectively, Drawn from Book-Level Ordered-Logit Models**



Notes. Each point represents a single book. Note that three outliers were removed to protect the scaling of the plot: two that had extremely high positive values on *ORDER* and one that had an extremely low value on *TIME*.

to which the effects vary materially across books, we have attempted to provide some characterization of this variance by estimating a series of book-level models of *STARS* as a function of both *ORDER* and *TIME*. We provide a plot of the book-level coefficients in Figure 2. It is critical to note that the key benefit of utilizing the panel-like approach described above is that we are able to pool the observations across books and reviewers—while controlling appropriately for heterogeneity—and thereby are able to increase the power with which we can identify these effects. Of course, we lose all of this power when we move to the book level, and thus many of the book-level dynamic coefficients are not significant. However, what Figure 2 does show is that the effects we see in Model (1) above do seem to vary across books. Therefore, even though the aggregate average effects are clear, any explanation for these dynamic factors must necessarily allow for the existence of both increasing and decreasing sequences of reviews. We return to these explanations below.

4.1. Robustness

A key alternative explanation is that there is, in truth, a single *TIME*-based process, although it is a nonlinear one. According to this argument, the *ORDER* variable may be serving as a proxy for this nonlinearity. Were this true, our inference that there are two dynamic processes would be erroneous. To check for this in as flexible a manner as possible, we reestimated Model (1) with a set of polynomial terms in *TIME*. Our primary concern was whether the impact of *ORDER* persists in these models or whether the inclusion of the nonlinear *TIME* terms rendered the information included in *ORDER* superfluous. We find that, in each of these specifications—up to a sixth-degree polynomial in *TIME*—*ORDER* is significant at the $p < 0.001$ level, as in Model (1). We conclude that *ORDER* does capture a separate and distinct dynamic process over and above a temporal one.[9] Although these results all point toward a consistent declining trend in *TIME*, we will challenge this conclusion in §4.2.

We assess the robustness of our results in a number of additional specifications. In Model (2), we estimate the joint dynamic effects on our entire data set of 74,750. This requires us to drop the *REVAVG* variable because we are only able to calculate reviewers' averages when they use a consistent identifier in their reviews. The consistency of these results with those in Model (1) suggests that the "pruning" required to calculate our reviewer-level measures does not drive our results.

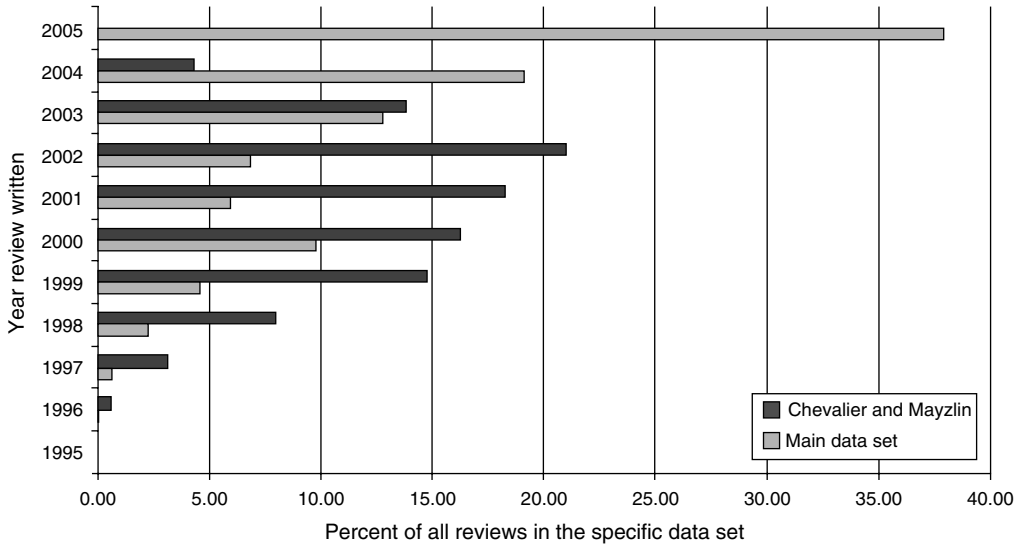
An important consideration in our analysis is the generalizability of our findings to a broader set of books. Although we have access to a data set that is, by most comparisons, quite large, these reviews are nonetheless drawn from a subset of 350 top-selling books. We stress again that we restrict our data gathering to a manageable number of books to ensure that our additional data collection efforts—in which we capture and use *all* of the other reviews for each of the reviewers in our data set (which generates a second data set of greater than 500,000 reviews)—is tractable. To assess the impact our sample selection may have on our characterization of the dynamic effects, we estimated an analog of Model (1) on a far broader data set. To do so, we obtained some of the main variables—*STARS*, a book identifier, and the date of the review—from the primary data set used by Chevalier and Mayzlin (2006). This data set consists of up to the first 500 Amazon.com reviews for 4,622 randomly chosen books.[10] Given the computational challenges inherent in estimating this many book fixed effects, we estimate instead a random-effects model for these books.[11] As shown in Model (3), we find

[9] These results are available from the authors upon request. It should be noted that some of these terms are significant in some of the models. In particular, in some of the cases, the quadratic and cubic terms are significant and positive. However, we stress that in none of the models does the explanatory power of the nonlinear *TIME* terms affect the significance or sign of the *ORDER* variable. Moreover, we stress that the impact of these nonlinear terms is small. In fact, even though both are positive, they never imply that the overall impact of *TIME* is negative in the range of the data set. Specifically, they imply that $(\partial/\partial TIME)U_{ib} < 0\ \forall\, TIME < 3{,}833$, which exceeds the maximum value in our data set.

[10] See Chevalier and Mayzlin (2006) for the specific randomization approach. Ours is a larger set of books than that analyzed by these authors because their final data set had to satisfy the criterion that the book was available—and had enough reviews—on BN.com as well. We are very thankful to Dina Mayzlin for providing these data for our analysis.

[11] We stress that these results should be interpreted with some caution given the possibility of correlation between the book effects

Figure 3 Proportion of All Reviews in Each Data Set That Were Generated in a Given Year: 1995–2005



consistent results on *ORDER*: there again appears to be a decline in ratings over the sequence. However, we find that this is not true of *TIME*. In this data set, although we are still able to infer the existence of two distinct dynamic processes, ratings seem to *increase* rather than decrease over time as they did in our focal data set. We use this contradiction to motivate a deeper inquiry into temporal dynamics in §4.2.

4.2. Investigating the Inconsistent Temporal Effects

In this section, we study in more depth the role of *TIME* in the evolution of ratings. Recall that the theory of temporal dynamics of Li and Hitt (2008) is based on a self-selection argument. In their model, buyers arrive, on average, in the order of their utility (expected and, ultimately, realized) for the product. Those who expect to like the product more buy it earlier than others. It is clear that (a) this would give rise to declining ratings; and (b) this is a temporal, as opposed to a sequential, theory: conditional on time lapsed, the number of submitted ratings would have no impact on the arrival time of various customer segments because it is the realization of an exogenous ordering of buyers, not the previous reviews themselves, that is driving the dynamic effect. In their paper, Li and Hitt present results that demonstrate the existence of the temporal phenomenon as well as data supporting the idea that consumers do not fully account for the dynamic bias in their decision making. It is important to note that these authors perform their analysis at the book level and are thus not able to control for important reviewer-level characteristics, which, as demonstrated in Table 4, play an important role in determining ratings. Moreover, as noted above, they do not control for *ORDER*. Having said that, even after controlling for these factors, Model (1) in Table 4 seems to corroborate their results. Model (3), however, does not.

To ascertain the underlying reason for these inconsistent results obtained on two different samples of books, we began by investigating the most salient differences between the data sets. One glaring difference is shown in Figure 3 and relates to the profile of the arrival dates of the reviews. This figure presents, for both our main data set and that used by Chevalier and Mayzlin (2006), the proportion of the data set's reviews that were generated in each year. As one would expect from a "top-seller" list, more recent books are more heavily represented in our data set than in the randomly chosen data set analyzed by Chevalier and Mayzlin (2006). Whereas their data are drawn from a relatively broad time period, ours is more focused on recent years. This raises the intriguing possibility that the different temporal results in Models (1) and (3) may be driven by an underlying dynamic trend in which more recent books are rated more negatively. To evaluate this explanation, we reestimate Model (1) and include dummy variables to capture the calendar year in which the review was written. The results are shown in Table 4, Model (4). In Model (5), we check again that our results are not driven by our pruning decisions by estimating

and the dynamic variables (for example, "better" books may have more reviews and thus possibly higher values of *ORDER*). It is for this reason that we have opted to estimate our main models with fixed effects for each book. We view this analysis as purely a robustness check.

Model (4) on our entire data set, again without the important reviewer-level term.[12]

Several interesting insights emerge from these models. First, it is clear that there is an underlying dynamic trend independent of any given book in which reviews written on Amazon.com in its earlier years (1997–2000) are reviewed more favorably—holding constant vertical quality—than are reviews written in more recent years. Most important for our analysis of temporal dynamics, we see that once we condition on this global time trend in our data set, *the sign on book-level elapsed TIME is reversed*. It appears that once the global underlying trend is removed—in particular, the relative negativity in years 2001–2005—ratings tend to *increase* with the length of time a book has been available on Amazon.com. This is now more consistent with our results in Model (3) in Table 4. In Model (6), we add the calendar-year dummies to our estimation using the data set from Chevalier and Mayzlin (2006) and note that it is qualitatively equivalent to that in Model (3). It is noteworthy that our results with respect to the sequential dynamic trend persist in all models in Table 4, demonstrating the robustness of our claim that there exist (at least) two distinct dynamic processes at work, one over *TIME* and another over *ORDER*. That more recent reviews are systematically more negative—holding constant book quality, reviewer effects, and time since the book's first review—is an interesting and, to our knowledge, novel finding. It is not entirely clear whether this is due to a general negative trend in attitudes or perhaps due to an effect specific to Amazon.com or the Web. Future research is needed to determine the extent to which this finding may extend to other domains: Have consumers, in general, become more critical? Moreover, additional research is required to rigorously ascertain the cause of the residual positive book-level time trend. We offer the conjecture that the arrival of new information—outside of online reviews—may be one explanatory factor. That is, over time, consumers speak with friends, read book critics' evaluations, and observe retailers' top-seller lists. All of this information—which is not captured in our data set and is not directly driven by reviews—should, ceteris paribus, result in people making better decisions and thus assigning higher ratings.

Having (i) established the existence of two dynamic processes and (ii) provided evidence that the temporal process is composed of both (positive) book-level and (negative) global dynamic effects, in the next section we turn to the development of a deeper understanding of sequential dynamics.

[12] Note that the year fixed effects are relative to the year 1997. We chose this because it was the first year with a nonnegligible number of reviews in our main data set.

5. Why Do We See Sequential Dynamics?

In a fully rational Bayesian model of consumer decision making, one would expect that, ceteris paribus, more information, such as the arrival of more reviews, would lead to better decisions and thus higher ratings. The first buyers of a product or service make their decisions more or less in the dark, whereas those that follow should benefit from the feedback from, and observation of, the early adopters. The results presented in Table 4 seem to contradict this view, however. They suggest that, for the average book and conditional on *TIME* and reviewer effects, the distribution of ratings is shifted downward as more reviews arrive for a given book. In this section, we attempt to shed more light on the underlying causes for this sequential dynamic process. Our approach will be to address first the existing explanations for the sequential dynamics (see §5.1) and then to offer and test two new theories (see §5.2), based on the expansion of purchase errors as a function of the arrival of reviews from dissimilar others. The critical distinction between sequential dynamics and temporal dynamics is that the former capture effects associated with the review environment itself, whereas the latter are driven instead by factors exogenous to the reviews such as selection based on expected value (Li and Hitt 2008). We hold constant these latter factors via our temporal controls (*TIME* and the review-year dummies) and investigate what impact the prior reviews themselves may have on a consumer's decision as to what rating to assign to a book of a given quality.

5.1. The Motivation to Post Reviews

One possible explanation for the sequential dynamics we observe in Table 4, offered by Wu and Huberman (2008), is that the existence of more previously submitted reviews implies that a new review needs to be more extreme—more different from the existing average—in order to have an impact. Assuming that the process of posting a review is costly, they argue, the reviews that ultimately get submitted (for positively rated books) will be lower and lower. Reviews "close to" the prevailing average will be submitted with low frequency because the reviewer will not deem the cost of submitting the post to be worthy of the benefit (the impact the review will have). The key results demonstrated by Wu and Huberman (2008)—using three distinct data sets—are as follows: (i) there is no polarization in a context in which the existing review environment is not observable (i.e., when one cannot assess a review's potential impact), (ii) claims that are inherently of high (low) quality are reinforced and become more positive (negative) as more votes are cast, (iii) ratings decline over the sequence in a univariate regression of *STARS* on *ORDER*, and

(iv) the expected deviation from the average rating $E[d_n] \equiv |STARS_n - 1/(n-1)\sum_{m=1}^{n-1} STARS_m|$—where $STARS_n$ is the nth review in the sequence—when plotted against n, in the aggregate, seems to be increasing.

Although offering a novel and intuitively appealing theory, as well as a number of interesting and varied analyses, the work of Wu and Huberman (2008) stops short of testing their theory directly.[13] Results (i)–(iv) primarily represent demonstrations of the decreasing ratings phenomenon and address only indirectly issues of costly reviewing and motivation. We note, in particular, the increasing deviation result noted above (iv). The authors' graphical presentation of the sequential trend of the deviation from the average is consistent with a declining ratings trend, to be sure. Thus, it provides additional support for the claim that the dynamic phenomenon exists. However, the result does not provide support for any specific theoretical explanation for why it would exist. Moreover, it is important to note that the increasing absolute-deviation result is also entirely consistent with an *increasing* ratings trend as well. Thus, this is a necessary but not sufficient condition for a declining sequential trend.[14]

We begin by revisiting the work of Wu and Huberman (2008). We do so for two reasons. First, with an eye toward discriminant validity, it is important that we demonstrate that our results go through while controlling for the Wu and Huberman (2008) primary drivers. Second, and perhaps more fundamentally, a rigorous test of the theory and its implications will be useful to our overall understanding of the phenomenon. Recall that the proposed mechanism in Wu and Huberman (2008) is based on the interplay between impact and cost and what this implies for one's decision to submit a review or not. We focus on identifying a proxy for the cost of submitting a review. With such a proxy, we can then test whether the sequential dynamic effect is more pronounced for higher-cost reviews than for lower-cost reviews. To see why this would be implied by their theory, imagine reviews are either high cost or low cost and that this discrete categorization is exogenous. When reviews are low cost, then even when the expected impact is low—perhaps because the rating is close to the prevailing average and/or there are already many other reviews—we would suspect that most reviews would nonetheless be submitted. As a result, the sequential decline hypothesized by Wu and Huberman (2008) should not be very strong; the theory should have no "bite." However, when the cost is high, then we would expect—according to the theory—that only those reviews that are expected to have a significant impact would be submitted. It would be for these reviews that we would expect the average declining sequential trend to be most pronounced. To test this, we propose using the variable *REVLEN*—the length of the written review, in number of characters—as a proxy for the effort required to submit a review.[15]

Specifically, we augment Model (4) in Table 4 by adding *REVLEN* as well as interactions between *REVLEN* and both *ORDER* and *TIME*. Our focal test of the theory will be on the *REVLEN* × *ORDER* interaction, which Wu and Huberman (2008) would predict to be negative: when the cost is higher, there should be a more pronounced declining sequential trend.[16] The results are presented in Model (1) in Table 5. There are several important insights. First, the results provide evidence for the theory: the negative coefficient on the *REVLEN* × *ORDER* interaction implies that, ceteris paribus, the decline in ratings as more reviews arrive is more pronounced when the review is longer, i.e., more costly. To our knowledge, this is the most direct test of—and strongest evidence for—the theory proposed by Wu and Huberman (2008).[17] Second, it is interesting as well to note that the main effect of *REVLEN* is negative and significant. This suggests that longer reviews are associated with lower ratings, all else being equal. This, on the surface, would seem to be somewhat inconsistent with previous research that has suggested an inverse U-shaped relationship between review length and rating (Chevalier and Mayzlin 2006). However, these previous results are based on summary statistics and not multivariate estimates, which hold other factors constant. Thus, our results should be interpreted as capturing the relationship between *REVLEN* and *STARS* conditional on,

[13] In a different setting, Zhang and Zhu (2011) demonstrate the general idea that people are concerned about the impact of what they post.

[14] As such, it important to note that our data also exhibit the same result: we find that the average deviation from the prevailing average is increasing over *ORDER*. In fact, we find that this is true holding constant all of the key covariates we use in the rest of the paper. This analysis is available from the authors upon request.

[15] Of course, for this to be a meaningful proxy for the cost of the review, we need to make the assumption that the reviewer knows ex ante what the review is likely to say and how much effort it will take to explain his or her position.

[16] We prefer this approach to an obvious alternative—performing a median split on *REVLEN* and estimating two different equations for several reasons. First, given the normalization one makes in estimating choice models of this form, it is typically considered inappropriate to make cross-model comparisons of coefficients. Second, such an approach would allow for two different sets of fixed and review-year effects (one for above and another for below the mean), which would add unwanted noise to our estimates. Finally, of course, a median split results in a possibly significant loss of information.

[17] The coefficient on *REVLEN* × *TIME* is marginally significant ($p = 0.072$).

Table 5 Sequential Dynamics

	Model (1)	Model (2)	Model (3)	Model (4)	Model (5)	Model (6)
	Wu and Huberman (2008) theory test		Does similarity decline over the sequence?	Similarity as a moderator and/or mediator of sequential dynamics		
	Ordered logit DV = *STARS*	Ordered logit DV = *STARS*	Linear model DV = *SIM30PLUS*	Ordered logit DV = *STARS*	Ordered logit DV = *STARS*	Ordered logit DV = *STARS*
REVAVG	6.56E–01***	6.54E–01***		6.50E–01***	6.48E–01***	6.51E–01***
	(1.49E–02)	(1.50E–02)		(1.61E–02)	(1.62E–02)	(1.50E–02)
ORDER	−1.29E–04***	−7.86E–05*	−9.93E–09*	−1.02E–04*	2.29E–05	−8.36E–05*
	(3.73E–05)	(3.85E–05)	(3.93E–09)	(5.29E–05)	(5.73E–05)	(3.85E–05)
TIME	2.36E–04*	3.56E–04**	−2.19E–08***	2.89E–04*	3.59E–04**	3.92E–04***
	(1.22E–04)	(1.23E–04)	(3.17E–09)	(1.25E–04)	(1.26E–04)	(1.23E–04)
CUMULMEAN		3.51E–01***			5.91E–01***	3.28E–01***
		(8.79E–02)			(1.60E–01)	(8.81E–02)
CUMULSD		−3.78E–01***			−4.08E–01*	−3.79E–01***
		9.92E–02			(1.84E–01)	(9.94E–02)
SIM30PLUS				2.31E+02	1.55E+02*	1.44E+02*
				(1.22E+02)	(7.47E+01)	(6.46E+01)
SIM30PLUS × *ORDER*				8.53E+00***	9.25E+00***	6.75E+00***
				(2.20E+00)	(2.16E+00)	(2.09E+01)
SIM30PLUS × *TIME*				−7.50E–01***	−7.66E–01***	−6.05E–01**
				(2.29E–01)	(2.22E–01)	(2.12E–01)
SIM30MINUS						2.91E+00
						(4.08E+01)
SIM30MINUS × *ORDER*						5.61E+00
						(3.77E+00)
SIM30MINUS × *TIME*						3.86E+00*
						(1.92E+02)
REVLEN	−4.98E–04***	−5.35E–04***		−6.74E–04***	−6.78E–04***	−5.81E–04***
	(8.98E–05)	9.10E–05		(1.06E–04)	(1.06E–04)	(9.13E–05)
REVLEN × *ORDER*	−3.34E–07***	−3.23E–07***		−2.66E–07**	−2.73E–07**	−3.19E–07***
	(7.81E–08)	(7.82E–08)		(8.41E–08)	(8.40E–08)	(7.82E–08)
REVLEN × *TIME*	−1.22E–07	−1.05E–07		−7.18E–08	−6.61E–08	−9.92E–08
	(6.79E–08)	(6.81E–08)		(7.33E–08)	(7.34E–08)	(6.82E–08)
Book effects	Fixed	Fixed	Fixed	Fixed	Fixed	Fixed
Rev year effects	YES	YES	NO	YES	YES	YES
N	34,747	34,368	30,353	30,044	30,044	34,368
LL	39,205	38,821	212,644	34,363	34,326	38,798
AIC	79,070	78,276	424,853	69,216	69,146	78,244

Notes. Robust standard errors are in parentheses. DV, dependent variable; LL, log likelihood; AIC, Akaike information criterion.
$^{*}p < 0.05$; $^{**}p < 0.01$; $^{***}p < 0.001$.

most importantly, accounting for the vertical quality of the book and the calendar date of review arrival.[18]

Another recent work (Moe and Trusov 2011) provides another set of relevant empirical results relating to sequential dynamics. In their book-level model of average ratings, they find that both the prevailing (i.e., lagged) mean and variance have an impact on the distribution of subsequent ratings. Thus, we reestimate Model (1) with additional covariates capturing the review environment:[19] the mean and standard deviation of previous ratings. It is important to note that there is little theoretical guidance to form hypotheses about the impact of these covariates; they can, and probably do, have a range of different impacts in different settings. With respect to mean, the theory pro-

[18] Indeed, our data set generates the same qualitative *REVLEN* − *STARS* relationship at the summary-statistic level as that reported by Chevalier and Mayzlin (2006). Specifically, average review lengths in our data are as follows: 173 (one star), 196 (two stars), 203 (three stars), 186 (four stars), and 147 (five stars). Moreover, to support the idea that it is important to control for other factors in analyzing review length, notice that the average review length is 107 characters for reviews before 2000 and 168 characters for reviews in 2000 or later.

[19] We have not included these covariates in our main model for two reasons. First, we are concerned about the possibility of introducing terms that are correlated with the error term. Specifically, this endogeneity may arise if a shock to book *b*'s ratings is correlated over time. Second, we are also concerned about the "reflection problem" (Manski 1993). That is, when modeling one's actions as a function of those of aggregate others, one risks misinterpreting the results as social effects when, in fact, they are possibly the result of correlated unobservables. Having said that, we see the results in Model (2) as an important demonstration of the robustness of our results.

posed by Wu and Huberman (2008) makes no clean prediction. However, as we show in Appendix C, we demonstrate that the context captured by Wu and Huberman (2008) may be consistent with the prediction that the impact of a marginal increase of the prevailing mean on the review distribution should, all else equal, be positive. One could also imagine that the prevailing average could establish a norm of sorts with respect to a book and thus might result in interreviewer herding effects. This would also imply a positive effect. The expected marginal impact of the spread of the distribution is equally unclear. On one hand, the impact of an increase in the spread on the marginal rating could be positive in that the widely diverging views assist readers in finding and making choices that better match their preferences. This would be consistent with Sun's (2010) theory. On the other hand, it might be the case that high variance would weaken the "norm" that a high mean might establish, granting license to a reviewer to post a more negative rating to a previously highly rated book.[20]

As shown in the results in Model (2) in Table 5, the qualitative nature of our previous results persists. Most notably, the marginal rating distribution is shifted downward as more ratings arrive and upward as more time passes. These results suggest that our primary results with respect to both temporal and sequential dynamics are not driven solely by the impact of the distribution of previous ratings. These factors do, however, appear to have some impact. The positive coefficient on the prevailing mean—conditional on vertical quality—would be consistent with both Wu and Huberman (2008), as well as herding behavior among reviewers. It is interesting to note the difference between this result and that shown by Moe and Trusov (2011), who find that the lagged mean is associated with fewer high ratings and more low ratings; this may be due to differences between the two categories (beauty products versus books),[21] as well as to the fact that we perform our estimates at different levels. The negative coefficient on the prevailing standard deviation implies that more noisy ratings are associated with lower subsequent ratings. This may be due to a weaker high ratings norm surrounding the review of a given book or perhaps to an increase in purchase errors, which we address in §5.2.

[20] Moe and Trusov (2011) estimate a hazard model of ratings as a function of lagged variance. They find that it increases both high and low ratings and thus the *spread* of the marginal rating. It is unclear from their analysis what the impact of this quantity would be on the *average* rating.

[21] As suggested by Moe and Schweidel (2012), there may exist significant heterogeneity with respect to people's reaction to the prevailing ratings environment. Thus, it may be the case that the relative prevalence of those preferring negative, as opposed to positive, environments is different across these product categories.

5.2. Purchase Errors and Reviewer (Dis)similarity

Given that we find in Models (1) and (2) in Table 5 that for *both high-cost and low-cost reviews* there is a declining sequential trend, it is highly likely that there is an additional mechanism (or set of mechanisms) behind this dynamic process beyond that based on motivation. In this section, we propose and test two related theories in an attempt to explain this decline. Because we have found support for the theory offered by Wu and Huberman (2008), we will control for the cost of posting the review via inclusion of the *REVLEN* term and its interactions. The two theories we propose here share a single common idea: as more and more reviews arrive, some people will benefit from the additional opinions and thus make better choices and assign higher ratings, all else being equal. Others, however, will make worse choices and, as a result, assign lower ratings. We present and test two theories that have the potential to explain why for some books it may be the case that additional reviews will lead to worse decisions. A key factor in determining whether additional reviews will help or not, we argue, is the extent to which they come from reviewers with *similar tastes*. We envision opinions about an object—a book, for example—to be composed of both signal and noise. Reviews from similar others, we propose, are valuable signals that help people to make better decisions. On the other hand, reviews from dissimilar others add noise to people's decision-making processes and may thus lead to worse decisions and lower ratings. The two theories we propose differ in the underlying mechanisms that may give rise to the negative association between ratings and *ORDER*. The first proposes that people systematically find it more difficult to identify reviews from similar others the more reviews that exist: it is harder to find "relevant" and diagnostic reviews in a set of 1,000 compared with a set of 100. The second theory proposes that, all else being equal, reviewers are systematically less similar to previous reviewers the later they arrive in the purchase queue. We subsequently discuss each of the theories at a qualitative level. We then provide in Appendix B a more thorough discussion of each within the context of simulation studies, which also provide evidence for the relationship between our observed pattern of ratings and the proposed mechanisms.

5.2.1. Theory 1: Errors in the Assessment of Diagnosticity. Imagine a simple two-step model of the decision-making process using online ratings. First, the reader assesses the quality of the information: the extent to which previous ratings reflect her preferences. Second, she then incorporates the ratings information into her beliefs about the product by weighting most heavily those reviews believed to be most diagnostic. We ask, how might this process be

affected by the arrival of more reviews? On one hand, more data are likely to have a positive impact on the second stage: more information should result in more accurate estimates of utility, conditional on accurate identification of the most relevant reviews. However, the impact of more reviews on the first stage is less clear. We propose that it is quite possible that it becomes more difficult to assess the diagnosticity of a large set of items compared with a small set of items, and thus, more errors are made. This theory is consistent with the rich marketing literature discussing and demonstrating the potentially negative impact of additional information on decision quality.[22]

One could imagine many possible behavioral mechanisms that would give rise to this prediction. As a particularly simple example, imagine that a prospective buyer allocates a fixed amount of time T to assessing the extent to which previous reviewers of a book have preferences that are similar to hers. Moreover, imagine that she uses the following simple allocation rule: allocate time T/N to each of the N reviews available on the product. This time would be spent reading the content of the review and perhaps the reviewer's other reviews. Then, as long as the accuracy of her assessment about review r is increasing in the time allocated to her consideration of this review, this process would give rise to more errors as more reviews arrived, all else being equal. These assessment errors would lead to more purchase errors, and as long as people negatively rate products that they expected to like but did not, this would give rise to declining ratings. Taken together, the overall impact of *ORDER* on ratings, according to this theory, will be the net result of the positive impact on the accuracy of utility assessments, conditional on weighting, and the negative effect on the diagnosticity assessment, or weighting. To the extent that the latter dominates, we would expect ratings to decline in *ORDER*. Otherwise, they would increase in *ORDER*. We emphasize that this fixed-time allocation process is simply an example of a process that would give rise to increasing errors as more information arrives. One could imagine many others of various levels of complexity.

We do not, of course, directly observe either the amount of time that a buyer spends assessing the relevance of reviews or which specific reviews she weights most heavily in her purchase decision. Thus, to test the theory, we again need to make use of proxies to identify when this theory will have explanatory power. We argue that this mechanism is likely to lead to more, and perhaps more pronounced, purchase errors the more heterogeneous are the reviewers over the sequence of reviews. For a given prospective buyer, the probability of a negative purchase error—in which the ex ante expected utility is high leading to a purchase but ex post utility is low—will be highest when (a) there are many preceding reviews and (b) her tastes are very different from those that posted reviews before her. In such a context, it is hard for a prospect to correct for the high level of heterogeneity. On the other hand, the prospect's reduced ability to decode the diagnosticity of the previous ratings will have less of an impact on purchase errors to the extent that the tastes of the prospect are more similar to the previous reviewers' tastes. To be clear, although we do not necessarily expect there to be fewer errors of this kind when reviewers are less heterogeneous, we simply expect the errors to have less of an impact on her experience and, thus, ratings.

Several additional comments are in order here. First, this explanation is very much specific to the horizontal quality context we are studying. Were quality purely vertical, there would be no reason for one to assess previous raters' preferences (stage 1), and thus, the impact on ratings of *ORDER* would always be positive. Second, one would expect no (or little) impact of offsetting errors of the positive type. That is, although it is theoretically just as likely that prospective buyers will not purchase books that they "should" purchase as the reviews pile up, these errors would not show up in the ratings data because nonpurchasers do not write reviews. Finally, we stress again that we propose here a behavioral mechanism and not one that would be consistent with a fully rational, forward-thinking actor. The marketing literature, however, has documented the fact that these behavioral forces can lead to significant declines in decision quality as a result of "information overload." We refer the interested reader to Appendix B, where we provide results from a simulation model supporting the idea that such a mechanism can lead to lower ratings over the sequence.

5.2.2. Theory 2: Increasing Dissimilarity. Our second theory of sequential dynamics also relates to the potential for purchase errors that increase in *ORDER*. The mechanism we study here is based on the evolving dissimilarity among buyers and, thus, reviewers. We propose that buyers may become increasingly less similar to the aggregate set of buyers that preceded them the more reviews that have arrived. As an example, imagine two prospective buyers, one arriving after five previous reviews and the other after 500 reviews. We suggest that it is likely that the former will, on average, have tastes that are more similar to the five that preceded her than would the latter with respect to the 500 that preceded her. As a result, on average, the ratings on which the former buyer bases her choices are likely to be more diagnostic of her expected utility than would be true

[22] See, for example, Jacoby et al. (1974a, b).

of the latter. In turn, then, her decisions are likely to be better ones, leading to, on average, higher ratings. Consistent with the view of "adopter categories" proposed by Rogers (1993),[23] we argue that adjacent adopters in the diffusion queue are similar to each other. This similarity may be due to any number of factors—from marketing to media usage to word of mouth. Simply speaking, we assume that buyer n is similar to buyer $n-1$, who in turn is similar to buyer $n-2$, and so on. Because this similarity is not perfect—n and $n-1$ are not carbon copies—it is reasonable to then assume that $n+1$ is "more similar" to n than she is to $n-1$. In the aggregate, then, this has the straightforward implication that any adopter n is more similar to other reviewers closer to her in the queue than to those that arrived in the more distant past. This, in turn, implies that aggregate measures—such as the mean rating—are more diagnostic of a prospective buyer's utility the earlier in the queue she arrives. To see this implication, note that those arriving later in the queue will view aggregate measures comprising opinions of both those that are similar to them and those that are dissimilar. The earlier buyers, on the other hand, will view aggregate measures that represent the opinions of others that are relatively more similar. Thus, these more diagnostic opinions should lead to better decisions and higher ratings.

For emphasis, we stress two important aspects of our theory. First, the mechanisms we propose should be viewed as complementary to the existing literature. In particular, our theory is not related to one's decision, conditional on purchase, as to whether or not to review. Analogously, the theory offered by Wu and Huberman (2008) is not at all related to similarity or purchase errors over the sequence. Second, we stress again that this is a behavioral explanation for the ratings-dynamics phenomenon. In a fully rational context, the buyer could integrate out the previous reviewers' tastes yielding a result that "more information is always better." However, as noted above, this view is inconsistent with the literature demonstrating potentially inferior decisions in higher-information contexts. We again note that Appendix B provides simulation studies in which we provide an example and demonstrate that it predicts declining similarity over the sequence of reviews, as we propose.

5.2.3. Testing the Theories. Both of our theories are closely related to the concept of *similarity* among the reviewers of a book. Theory 1 suggests that a sequential ratings decline can occur when more reviews lead to errors in the assessment of diagnosticity, which are most likely to affect purchase-decision quality when reviewers are highly dissimilar. Theory 2, on the other hand, is based on the idea that dissimilarity itself increases over the sequence of reviews. In both theories, we hypothesize that increasing purchase errors lead to lower ratings. To test and identify distinctly these two proposed mechanisms, we exploit the fact that similarity should impact ratings differently conditional on each of the theories. Theory 1 implies a *moderating* role of similarity on the relationship between *ORDER* and ratings: the impact of one's declining ability to assess diagnosticity should be more pronounced when previous reviewers are more dissimilar. On the other hand, Theory 2 implies a *mediating* role for similarity: the increase in *ORDER* (the arrival of more reviews) *leads to* higher dissimilarity, which in turn drives down ratings. We will thus test for both the mediating and moderating role of similarity.[24]

Next, we describe how we construct a measure to capture the similarity among reviewers in terms of the set of books they choose to read and review.[25] We first define the following:

$$B_r \equiv \text{set of books reviewed by } r,$$

$$S_{br}(t) \equiv \text{set of reviewers who reviewed book } b \text{ at least } t \text{ days prior to reviewer } r.$$

On a pairwise basis, the empirical probability that a book reviewed by r has also been reviewed by s is calculated as follows:

$$M_1(r,s) \equiv \frac{N(B_r \cap B_s) - 1}{(N(B_r) - 1)},$$

where $N(W)$ is the cardinality of set W. We subtract 1 from both the numerator and denominator to remove the impact of the focal review. That is, for book b in our data set, we want to know how similar reviewer r is to a previous reviewer s of the same book b. To do so, we look at all reviews on books *other than* b. The numerator of M_1 is a count of the number of other reviews the two reviewers have in common, whereas the denominator controls for the volume of other reviews generated by r.

Whereas M_1 is a pairwise similarity measure, we want to construct a measure of the aggregate similarity between r and *all* s that preceded her. To do so, we take an average of the values of $M_1(r,s)$ over all previous reviewers s for book b while also adding one more normalization to account for differences

[23] See, for example, Figure 7–2 of Rogers (1993, p. 262).

[24] We thank Jonah Berger for first suggesting this distinction.

[25] We have also analyzed a similarity measure that focused instead on the similarity between reviewers' *ratings* on books for which they both posted reviews. Because the explanatory power of this measure was far lower than that of the measure we utilize here, these results are not presented.

across the preceding reviewers in terms of how many reviews they generate; we define it as follows:[26]

$$SIM(r, b, t) \equiv \frac{1}{N(S_{br}(t))} \sum_{s \in S_{br}(t)} \frac{M_1(r, s)}{N(B_s) - 1}. \qquad (3)$$

Thus, again, the numerator reflects the fact that some j may have contributed hundreds of other reviews—and would thus be very likely to have a high overlap with i—whereas others may have contributed few other reviews. Our focal measure will be $SIM30PLUS \equiv SIM(r, b, 30)$. That is, we calculate the similarity of reviewer r to those reviewers that arrived at least 30 days prior. We do so because our theories are based on the idea that r's rating will be driven by her satisfaction with her purchase, which she made as a function of s's review. This lag accounts for the delay between s's influence on her purchase and her reporting of her satisfaction via her review. Thus, we assume that, on average, it will take 30 days from the time r reads s's review and provides her own review of the purchase that followed.[27]

The mean of *SIM30PLUS* in our data set is 1.22E−5 with a standard deviation of 2.16E−4.[28] We first test our assumption that similarity declines in *ORDER*. As shown in Table 5, Model (3), consistent with the motivation behind Theory 2, *SIM30PLUS* is decreasing in both *ORDER* and *TIME*.[29] Moreover, we note that the book fixed effects in this model add significant information to the model.[30] This suggests that there are important interbook differences in terms of the extent to which their readers are similar to each other. To the extent that we demonstrate that similarity plays a role in the average sequential dynamics we observe, this book-level heterogeneity in terms of similarity may offer an explanation for the heterogeneity we observe in these dynamic effects (see Figure 2). Note, however, that these results need to be interpreted with caution because the overall fit of Model (3) is quite low. Nonetheless, to our knowledge, this aspect of diffusion—the dynamics of interadopter similarity—has not been noted in the literature. We believe that this represents a potentially fertile area for future research.

The primary empirical test of our theories is accomplished by including *SIM30PLUS* in our model both as a main effect and as an interaction with *ORDER* and *TIME*. The moderating role of similarity (Theory 1) is thus evaluated by inspection of the coefficient on its interaction with *ORDER*. The mediating role of similarity (Theory 2) will be evaluated by comparing the coefficient on *ORDER* across specifications with and without *SIM30PLUS*. We again estimate the ordered-logit model via maximum likelihood. The results are shown in Table 5, Models (4)–(6). We begin with Model (4), which does not include the prevailing environment covariates. Note the significant moderating effect of *SIM30PLUS* on the *ORDER–RATINGS* relationship, as captured in the positive coefficient on the *SIM30PLUS* × *ORDER* interaction: the sequential decline in ratings seems to be most pronounced when the similarity of a reviewer with previous reviewers is low. This provides support for the diagnosticity assessment theory (Theory 1). We stress that this result is not contingent on any particular value of *REVLEN*: it is not the case that we see the decline in *RATINGS* in *ORDER* only when the review is a high-cost one. We interpret this as evidence that our theory relates to a process distinct from that proposed and studied by Wu and Huberman (2008). Note that the main effect of *SIM30PLUS* was marginally significant ($p = 0.059$); this might suggest that even for early reviews, higher similarity among reviewers seems to be associated with higher ratings, suggesting a higher degree of satisfaction with the purchase. In Model (5), we include the covariates *CUMULMEAN* and *CUMULSD* to demonstrate the robustness of the results. Again, we find evidence for the moderating role of similarity (Theory 1). It is useful to note that the inclusion of *CUMULSD* in particular helps to exclude a possible alternative explanation for the significant effects on *SIM30PLUS* × *ORDER*: similarity may be a proxy for the noise in the previous ratings that may lead to more purchase errors. That is, one might expect that more variance in ratings leads to more purchase errors as it makes it more difficult to come to a clear purchase decision. Furthermore, we might expect that this variance would be correlated with the heterogeneity among the previous reviewers. By including *CUMULSD* in Model (5), we are thus holding constant this variance and demonstrating that, in fact, *SIM30PLUS* seems to be capturing an element of

[26] To understand the justification for this second normalization, consider the following example: $B_1 = \{X, Y\}$, $B_2 = \{X, Y\}$, and $B_3 = \{X, Y, L, M, N\}$. The question is whether reviewer 1 is equally similar to 2 and 3 or more similar to 2. We take the view that the latter is more likely, which we accomplish by this second normalization. We are grateful to an anonymous referee for suggesting this approach.

[27] We are grateful to an anonymous referee for suggesting this approach to lagging the impact of the review on the prospective purchaser's opinions.

[28] Recall that the measure is normalized by the product of the number of reviews that both reviewers produced in total.

[29] The results in Model (3) in Table 5 also suggest that a reviewer of a given *ORDER* will be less similar to the *ORDER* − 1 reviewers that preceded her the more time that has elapsed since the book has been reviewed on Amazon.com. Although we did not offer a hypothesis on this quantity, we offer the following conjecture: it is possible that the fact that it takes a book longer to generate the *ORDER*th review may suggest that the information about the book is traversing more weak ties—which are slower in sharing information—as it moves from network to network. The fact that there thus may be more inter (compared with intra)-network recommendations would certainly be consistent with a lower degree of interreviewer similarity.

[30] As measured by a comparison between the Akaike information criterion for Model (3) and another model without the fixed effects (not shown).

heterogeneity among buyers and reviewers that is not fully captured in their observed ratings.[31]

We are not able to test cleanly in Models (3) and (4) for the mediating role of similarity (Theory 2). This is because the sample over which we estimate these models—compared with Models (1) and (2)—is different: we exclude those reviews that arrived in the previous 30 days. To test Theory 2, we need to use an identical sample in order to compare the coefficient once the similarity variables are included. Because *SIM30PLUS* is not defined for the most recent 30 days of reviews, we create another similarity term, *SIM30MINUS*, which is the complement of *SIM30PLUS* in that it captures the similarity of the reviewer to the authors of the reviews that arrived within the previous 30 days. We then include this variable—along with its interactions with *ORDER* and *TIME*—in our model and estimate it on the entire sample used in Models (1) and (2). The results are shown in Model (6).[32] In evaluating Theory 2, the relevant comparison is on the *ORDER* variable between Models (6) and (2). Because we see no attenuation of this coefficient, we reject Theory 2.

To summarize our analysis of sequential dynamics, we have established that the theory proposed by Wu and Huberman (2008) seems to explain some, but not all, of the variance in our data set. We have also found strong support for a theory based on the decreasing ability of readers to assess previous reviews. As more reviews arrive, it is harder to make use of them, leading to more purchase errors, on average. Finally, we found no support for an alternative theory based on declining similarity among reviewers. We stress again that the results we demonstrate on the impact of dissimilarity are derived while controlling for the cost of the submitted review. Because no existing theory can explain the results we find with respect to the impact of *SIM30PLUS*, we see our theory as distinct from, and complementary to, those currently offered in the literature (Li and Hitt 2008, Wu and Huberman 2008).

6. Conclusions, Limitations, and Directions for Future Research

We began with the observation that, on average, ratings for books decline in both *TIME* and *ORDER*. We established as the primary objective of the paper the development of a deeper understanding of the causes behind this phenomenon. In the course of this analysis, we have both provided additional tests of extant theories (Wu and Huberman 2008) and generated and tested additional theories. We see five primary contributions of the paper.

(1) *Multiple, distinct dynamic processes occurring simultaneously:* As demonstrated in Table 4, the impact of each of the dynamic variables, *TIME* and *ORDER*, is distinct. The passage of time has one effect (or set of effects), whereas the arrival of more reviews has another.

(2) *Diagnosticity assessment as a driver of declining sequential trend:* We found that the similarity between a reviewer and those that preceded her in the sequence moderates the impact of *ORDER*—the number of reviews that have already arrived—on the ratings distribution. Specifically, although this impact is negative on average, it is most pronounced when this similarity is low. We interpret this as evidence for our theory that the sequential decline is partially due to the increasing difficulty one may face in determining which reviews are relevant to one's choice and, in turn, implementing that choice. Moreover, the result demonstrates why some books may experience an increasing trend—those with more homogeneous adopters—whereas others (and, indeed, on average) experience a sequential decline.

(3) *Temporal dynamics driven, in part, by a broader trend:* The observed temporal decline in ratings seems to be associated with calendar time and not necessarily the amount of time a book has been available for review. In fact, once this calendar trend is controlled for, we find that the residual dynamics are *increasing* rather than decreasing. Because this result is independent of any specific book, it suggests that there may be an overall economy-wide drop in consumers' ratings of products. Whether this is driven by experience with the "ratings process," a reaction to a broader set of others' ratings, the macroeconomic environment, or some other factor(s) represents a potentially fruitful area for future research. It is important to stress that these results do not necessarily represent a disconfirmation of Li and Hitt's (2008) theory. It is certainly possible that, for some books, the self-selection dynamic can still be demonstrated while controlling for this important covariate. Future research and perhaps richer models are needed to establish this. We address this further below.

(4) *Evidence for motivation-based theory:* Our analysis using the length of the review as a proxy for "cost" would seem to provide support for the theory offered by Wu and Huberman (2008). Specifically, our results showing that the sequential decline in ratings is more pronounced for high- than for low-cost reviews are consistent with their theory. We also found a positive impact of the prevailing mean on the ratings distribution, which may be seen as consistent with their theory as well. Of course, given this,

[31] We also estimated a model in which we interact *CUMULSD* with both *TIME* and *ORDER*. The qualitative nature of these results with respect to the variables of interest is unchanged.

[32] For reasons of exposition, we do not present the results of an analog of Model (6) that does not include *CUMULATIVE AVG* and *CUMULSD*. These results are qualitatively equivalent to those in Model (6) and are available from the authors upon request.

it was important that our subsequent analysis of the sequential process controlled for the review cost and, thus, their motivation-based theory. Our subsequent results demonstrate that our theory is complementary to theirs in that the effects we and Wu and Huberman predict appear to exist simultaneously and independently from each other.

(5) *Declining similarity:* Although not a focal element of our paper, we also demonstrated that interreviewer similarity seems to decline over the sequence of reviewers. We offer a "quasi-random walk" model in Appendix B that explains such a sequence, and then we demonstrate that the pattern is consistent with our data. Other than Rogers (1993), we know of no research that has directly addressed the evolving profile of adopters. However, we see it as an interesting area for future research. To the extent that the diagnosticity of a set of reviews is declining the more distant that review is in the arrival sequence, it would suggest that the presentation of reviewer ratings should optimally account for this. For example, one might imagine that an average rating could be weighted to reflect more recent reviews given their higher usefulness to current prospective buyers.

These results have important implications for both researchers and managers. With respect to the former, we would suggest that future researchers interested in studying the dynamics of opinions, particularly online opinions, must consider the underlying dynamic trend outside of any individual item's ratings. Our results in §4.2 make abundantly clear that researchers studying the temporal dynamics of online reviews must control for calendar time, particularly to the extent that their data set spans a number of years, as ours does here. Without doing so, our results suggest that incorrect inferences may be drawn. To our knowledge, it is not the norm in the current literature to control for such factors. Moreover, our results are quite clear in demonstrating the existence of separate and distinct temporal and sequential processes. Thus, researchers interested in studying just one of these must necessarily control for the other.

Our paper has highlighted a range of interesting questions that researchers may pursue in the future. For example, what is the cause of this underlying calendar-year decline in ratings? Moreover, how general is it? Does this also occur in other domains? In addition, we have provided some evidence for a decline in similarity along the sequence of adopters. Again, is this general? Can we see this in other domains? Is this both an on- and off-line phenomenon? Furthermore, our diagnosticity assessment theory would suggest that some readers may make worse choices as more reviews arrive. This implies a range of interesting and important research questions surrounding the ways in which better formats could be designed to match buyers to the appropriate reviews. Finally, from a methodological standpoint, the next step in this research stream would be to specify and estimate a model that captures at a deeper level the extent and impact of heterogeneity across both books and reviewers. Regarding the latter point, we see the development of richer models to account explicitly for the self-selection that Li and Hitt (2008) predict as potentially useful. For example, one could imagine a Bayesian hierarchical model in which the data are augmented with the inferred utilities of sequential adopters. This might allow for a more direct test of the theory.

The implications our results have for managers are also quite important. On one hand, it is clear that the way in which consumers use online reviews will have important implications for both their purchases and the future reviews they submit. Our results on sequential dynamics, in particular, suggest that too many reviews may lead to information overload. Managers need to consider various ways to counter this potential hazard. The provision of simple aggregate summary statistics—means, histograms—is unlikely to solve the problem, which has its root, we claim, in the match between a prospective buyer's tastes and those of previous reviewers. Anything that can help to establish the quality of that match has the potential to lead to more satisfied customers.

Broadly speaking, our results suggest that the adoption by a more heterogeneous set of customers may lead to a decrease in dissatisfaction, both marginal and average. This, we argue, is a function of the diffusion process and reliance on online reviews. Moreover, it occurs even when quality is static, as would be the case for books. This leads to two implications for managers. First, the results alert practitioners to the fact that their reviews may exhibit a declining pattern for no reason other than that it was adopted by a relatively dissimilar set of customers. That is, decreasing reviews may not always imply a change in quality, either actual or perceived. In fact, this is reinforced by the calendar-date result for temporal dynamics, which represents yet another quality- and competition-independent source of ratings decline that should be controlled for in any business analysis using as an input a dynamic data set of online reviews.

Perhaps more interesting, the result implies that firms may want to take into account the potentially adverse impact of heterogeneous diffusion in their marketing—especially word-of-mouth marketing—implementation. This is particularly noteworthy in that the typical prescription—following the sociology literature (Granovetter 1973, Burt 1992)—is to foster the spread of information across highly different

groups of adopters. Our results suggest that a potential caveat to this approach is that it may lead to more purchase errors, lower satisfaction, and, perhaps, long-term sales declines.

Of course, our paper is not without limitations. Primary among these is the fact that we have focused on *average* book effects, although there is likely to be strong interbook heterogeneity. Specifically, although we control for heterogeneity in vertical quality via book fixed effects, we specify dynamics as book independent. Thus, it is imperative that the reader interpret our results as demonstrating an *average* sequential decline and as providing an explanation that seems to hold *on average*. Our results suggest that similarity among reviewers is a key moderator: when similarity is high, we do not expect these declines to occur. However, we would expect that there may exist other important moderators. Future research would be needed to identify and explain these other sources of book (or more generally, item)-specific dynamics: Do "better" books decline more slowly or, perhaps, increase? Do books from first-time authors evolve differently?

A second important limitation of our study is that it has focused exclusively on a single category: books. Whereas this is true of Li and Hitt (2008) as well, Wu and Huberman (2008), on the other hand, studied three different data sources. Our theories are in no way category dependent: we would expect that the same impact of similarity and diagnosticity would hold in other categories including, for example, restaurants or movies. However, future research would be needed to establish this. Surely, the fact that Moe and Trusov (2011) report different results on some variables suggests the need to conduct studies of this kind across multiple categories.

Finally, an important limitation of our analysis is that although we offer the first analysis that controls for the reviewer's motivation to post, this control is far from perfect. On one hand, the identification of proxies for the cost of posting a review other than *REVLEN* would be valuable. Given that it may be difficult to identify such a proxy, we suggest that an experimental test of this theory is likely to be a fruitful direction for study. For example, one could manipulate the cost of posting a review in any number of ways. One could also limit the number of reviews subjects are allowed to write, where the limit could be manipulated between a low number (high cost) and a high number (low cost). One might also manipulate cost more implicitly, perhaps by increasing or decreasing the "hassle" associated with the logistics of submitting the review: the number of clicks one needs to implement, administrative overhead associated with a review, etc.

On the other hand, we also fully acknowledge that there are aspects of motivation that will go well beyond how costly it will be to say what one wants to say. Future research will be needed to expose both our theories and those of Li and Hitt (2008) and Wu and Huberman (2008) to a set of tests that more completely reflect the potential impact of motivation. We would suggest that it is also likely that an experimental approach may be suitable for such an investigation into the role of impact. It would be relatively straightforward, for example, to manipulate the review environment in such a way that the impact of a review is likely to be high or low (for example, if there are few, versus many, existing reviews). The attractiveness of this setup would be the ability to control perfectly for the quality of the underlying experience. With respect to motivation, we note two recent papers that begin to address this issue: Hu et al. (2009) and Moe and Schweidel (2012). The former model, in a static context, the differential motivations to submit a high rating versus a low rating. Similarly, in a dynamic context, the latter study the interreviewer differences in reviewing motivation. They demonstrate that the review environment—what reviews have already been submitted—may lead to an "adjustment effect" that affects what rating one posts, holding constant the true underlying utility. By allowing for different assumptions on the arrival rates of the different customer segments, they show that this heterogeneity may also be able to explain the declining (sequential) patterns.

Finally, as is the case for all current research in this domain, we are limited in terms of our understanding of the relative attention paid by reviewers to various aspects of the prevailing review environment. We assume here that reviewers attempt to decode the diagnosticity of previous reviews and reviewers. Other researchers (Wu and Huberman 2008, Moe and Trusov 2011) assume that reviewers pay attention to summary statistics. However, we know of no study that investigates this question either empirically or experimentally. In this regard, we are excited about the prospect of new research technologies—perhaps eye tracking, for example—being applied to address these questions: What is the relative weight given to summary statistics versus individual reviews? Within the latter, how do people allocate their attention to some specific reviews over others? The answers to these and related questions would help a great deal in extending our understanding of both the usage and impact of reviews, online and otherwise.

Acknowledgments

The authors thank Dina Mayzlin, Monic Sun, Feng Zhu, and participants at the following conferences for their comments on previous versions of the paper: Marketing in Israel V, the Summer Institute on Competitive Strategy (Berkeley), the Wharton Interactive Media Initiative, and the Yale Center for Consumer Insights. The authors also express their thanks to Dan Ariely for his encouragement and support.

Appendix A. Summary of References Related to the Present Study

Reference	Research context	Dynamic variable(s)	Level of analysis	Empirical results on opinion dynamics	Theories offered	Evidence	Control for				
							Unit-level effects	Individual-level effects	Sequence	Elapsed time	Calendar time
Li and Hitt (2008)	Books	Time	Unit	Decline	Self-selection at purchase	Evidence for phenomenon but not theory	Yes	No	No	Yes	No
Wu and Huberman (2008)	Books, votes	Sequence	Unit	Decline	Motivation	Evidence for phenomenon but not theory	No	No	Yes	No	No
Moe and Trusov (2011)	Bath, fragrance, and beauty products	Sequence	Unit	Decline	Self-selection	Evidence for phenomenon but not theory	Yes	No	Yes	No	No
Moon et al. (2010)	Movies	Sequence	Rating	Decline at unit (movie) level	Self-selection	Evidence for phenomenon but not theory	Yes	Yes	Yes	No	No
This paper	Books	Sequence, calendar time, and book-level elapsed time	Rating	Decline over calendar time, increase over book-level elapsed time, and decline (increase) over sequence when similarity with previous reviewers is low (high)	Diagnosticity assessment	Supporting evidence for (i) motivation-based theory, (ii) moderating role of diagnosticity assessment, and (iii) declining macro-temporal trend No support for (i) self-selection and (ii) mediating role of diagnosticity assessment	Yes	Yes	Yes	Yes	Yes

Appendix B. Simulation-Based Models of Theories 1 and 2

Theory 1: Errors in the Assessment of Diagnosticity

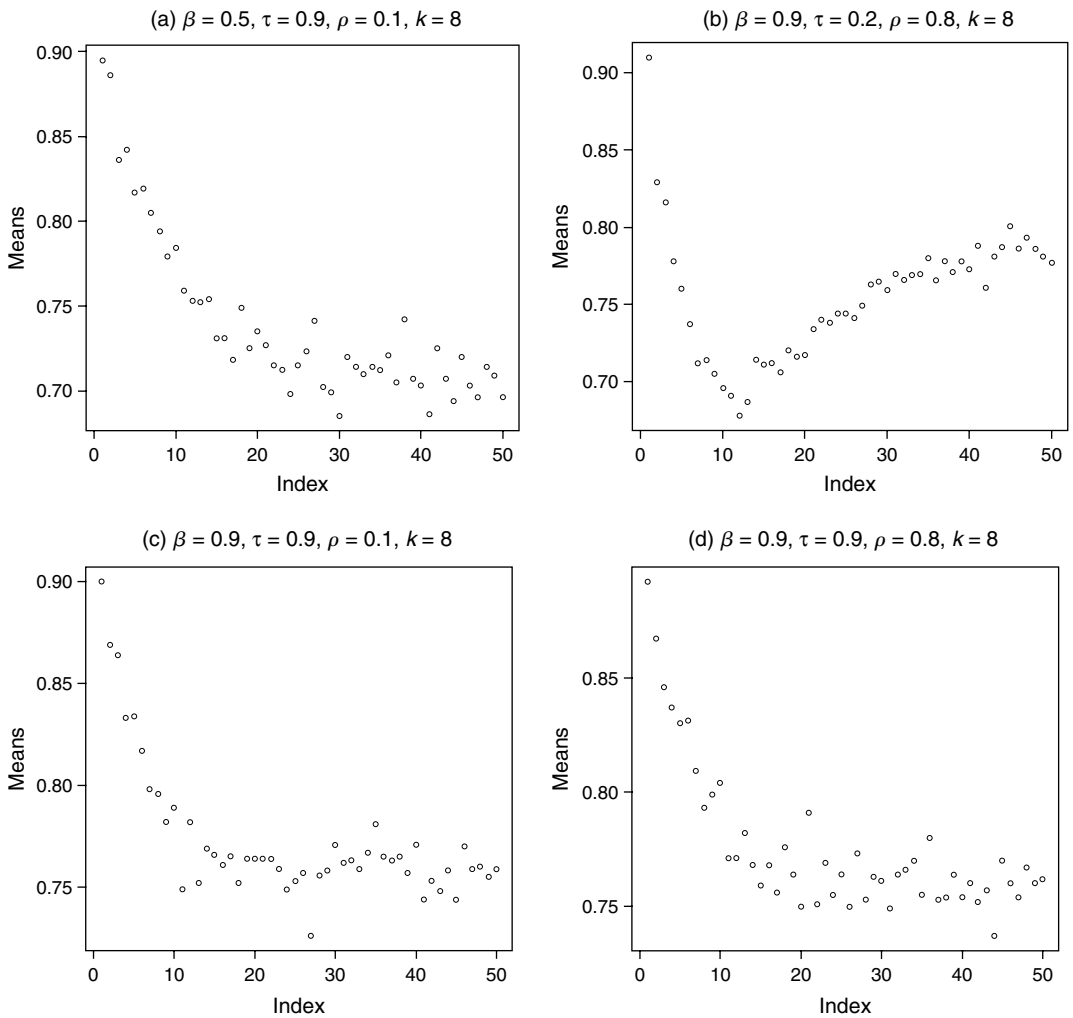
In this appendix, we provide a simulation-based model that demonstrates that a simple yet reasonable set of behavioral assumptions can lead to the conclusion that more available reviews leads to more purchase errors, which lead to lower ratings. The following are our primary assumptions.

Consumers randomly arrive as prospective buyers for a book b. Each prospective buyer is randomly assigned one of $n \in N$ types. Each type is randomly and independently assigned a $\theta_{nb} \in \{0, 1\}$ such that type n either "likes" ($\theta_{nb} = 1$) or dislikes ($\theta_{nb} = 0$) book b. We assume that $\Pr[\theta_{nb} = 1] = q_n$. Because these assignments are independent, only information from other buyers of the same type is helpful to one's decision making. A proportion $\beta \in \{0, 1\}$ of prospective buyers read the reviews, whereas $1 - \beta$ choose on their own without reading reviews. When choosing on their own, these customers make the "right" decision—buying a book when $\theta_{nb} = 1$ and not when $\theta_{nb} = 0$—with probability p. When the β buyers face a context without reviews, they do not buy. If there are reviews, they buy as long as the proportion of those reviews they deem to be relevant exceeds some threshold level τ.

These assumptions are meant to capture the setting modeled by Van Den Bulte and Joshi (2007), in which there are some people who are best described as "imitators" because they follow the decisions of others. Other people, however, can best be thought of as "innovators" because they make their own decisions. If one buys a book, regardless of the information that went into the decision, their true preference θ_{nb} is revealed as a function of their type. They then decide whether to write a review, which they do with probability ρ.

Facing a set of reviews, the β buyers attempt to determine which among them are written by people who are of the same type as they are. The assumption we make, consis-

Figure B.1 **Simulation Results: Theory 1**



Notes. Each panel represents a different set of parameters. Common across the four models is the fact the model was run until 50 reviews were generated. The results presented reflect the review on the x axis (1 to 50) and the proportion of reviewers who liked the book they bought on the y axis. These results, for each model, reflect the average of 1,000 iterations of the model at the specified parameter values. Also common across each model was the assumption that $p = 0.9$ and $q_n = 0.6\ \forall n$.

tent with the information overload literature in marketing (Jacoby et al. 1974a, b), is that the probability π that they are able to classify correctly a review as "diagnostic" (that is, written by someone of the same type n) or not is decreasing in the number of reviews the buyer must read. We adopt the following flexible specification:

$$\pi \equiv \Pr[\text{correct classification} \mid M \text{ reviews}] = \frac{1}{\sqrt[k]{M}}.$$

Therefore, the probability that she correctly classifies each of the M reviews is given by π. The realization of this classification is drawn iid for each review. The parameter k allows us to control how quickly the probability declines in M: higher values of k correspond to more gradual declines.[33]

In this simple setup, buyers can only report whether they liked or disliked the book they purchased, although they need not report anything. As Figure B.1 demonstrates, under a range of very different parameter values, we find that the average rating—or, alternatively, the probability that previous reviewers liked the book—is declining over all or part of the sequence of reviews.

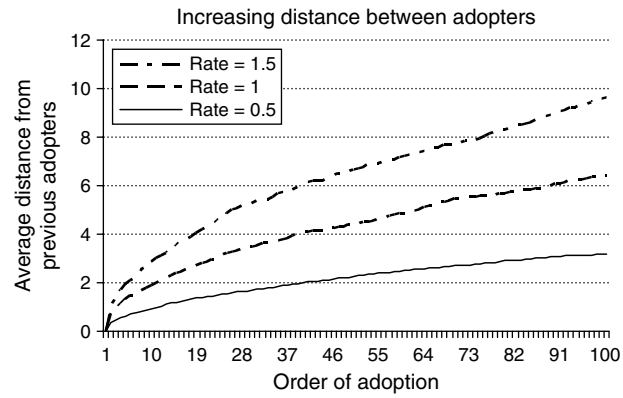
We stress particularly the fact that ρ, the probability that one writes a review conditional on buying the book, does not seem to have a particularly strong impact on the results. This can be seen in a comparison of panels (c) and (d), which present a context in which only 10% of the buyers write reviews and 80% of the buyers write reviews, respectively.

Theory 2: Increasing Dissimilarity

The critical assumption behind our increasing dissimilarity theory (Theory 2)—distinct from that of Theory 1—is that similarity may decrease exogenously over the sequence; the nth buyer is more similar to the set of $n-1$ buyers that preceded her than is buyer $n+1$ to the n that preceded her. As a result of this increasing dissimilarity, later buyers make more errors—the reviews on which they base their decisions are less predictive of their utility—and thus ratings decline. We demonstrate that declining similarity is the result of a very simple and plausible taste-space model. Imagine that tastes can be characterized by a two-dimensional space. The first adopter that arrives is centered at the origin, without loss of generality. Adopters $n > 1$ arrive according to the following process: we draw a circle of radius r around the taste location of adopter $n-1$, where r is drawn from an exponential distribution of mean λ. We then select adopter n's location on this circle from a uniform distribution. Thus, intuitively, adopters evolve in taste space such that they are a common average distance away from the previous adopter but that distance can be represented by any possible combination of tastes consistent with that distance. Therefore, this evolution is consistent on average but characterized by a higher degree of stochasticity.

What we are interested in is the similarity between adopter n and the $n-1$ adopters before her. To measure this, we calculate the average Cartesian distance from adopter n to each of the previous $n-1$. As shown in Figure B.2,[34]

Figure B.2 **Simulation Results: Theory 2**



the average dissimilarity (distance in taste space) increases strictly over the sequence, as expected. Moreover, the rate at which this increases is increasing in the mean of the exponential distribution. This is not surprising: the more different reviewer n is from reviewer $n-1$, the more different she will be on average from all of those who preceded $n-1$.

Appendix C. Implication of Wu and Huberman (2008)

Proposition. *Given Wu and Huberman's* (2008) *theory, the impact of an increase in the prevailing average rating among the $n-1$ ratings will be dependent on the distribution for the nth rating. The following are two sets of sufficient conditions for this impact to be positive*: (a) *additional positive ratings will not yield a sufficient impact to outweigh their cost*; *or* (b) *the distribution of the nth rating is J-shaped, and positive and negative ratings are equally impactful.*

Proof. Let Δ_n be the difference between the nth reviewer's proposed rating and the average of the $n-1$ posted ratings:

$$\begin{aligned}\Delta_n &\equiv R_n - \frac{1}{n-1}\sum_{m<n} R_m \\ &\equiv R_n - M_n,\end{aligned}$$

where R_m is reviewer m's rating and M_n is the prevailing mean when reviewer n "arrives." Of course, reviewer n may not post his rating. He will only do so if the "impact" of his rating will outweigh the cost. We will define the impact as $I(\Delta)$ such that $I(\cdot)$ is increasing in the absolute value of Δ, consistent with Wu and Huberman (2008). Moreover, we will define the cost if posting the rating as C, which we will hold constant in this analysis. Thus, she will post her review if and only if $I(\Delta_n) < C$. Now, define the pair $\{\underline{\Delta}_n, \bar{\Delta}_n\}$[35] as the "indifference deviations" such that

$$I(\underline{\Delta}_n) = I(\bar{\Delta}_n) = C,$$

[33] At $k=1$, the probability declines from 1 to 0.029 from the 1st to the 35th review. With $k=9$, on the other hand, it declines from 1 to 0.674.

[34] For each value of λ, the mean distance between adopters n and $n-1$, we run 1,000 iterations of each model with 100 adopters each.

[35] Two notes are warranted: (i) To be more precise, we would define these cutoffs as a function of C. Because we are assuming this is constant here, we ignore this for the sake of exposition. (ii) There is no reason to expect symmetry here. That is, it need not be the case that $\underline{\Delta}_n = -\bar{\Delta}_n$ or that $I(x) = I(-x)$.

where $\underline{\Delta}_n$ ($\bar{\Delta}_n$) is defined as the minimum difference between reviewer n's proposed rating and the prevailing average such that she posts the rating. This gives rise to a posting decision, whether she posts the review or not, $\pi = \{0, 1\}$, such that

$$\pi = 1 \;\Leftrightarrow\; \Delta_n < \underline{\Delta}_n \quad \text{or} \quad \Delta_n > \bar{\Delta}_n,$$
$$\pi = 0 \;\Leftrightarrow\; \Delta_n \in [\underline{\Delta}_n \bar{\Delta}_n].$$

Now, if the proposed rating R_n is distributed on $(\underline{R}, \bar{R})$ according to the cumulative distribution function $F(R)$, then we can easily calculate the expected value of rating R_n as a function of the decision to rate or not as specified above:

$$\begin{aligned} E[R_n] &= \int_{\underline{R}}^{\bar{R}} 1(\pi = 1 \mid \Delta_n, C) R_n \, dF \\ &= \int_{\underline{R}}^{M_n} 1(\pi = 1 \mid \Delta_n, C) R_n \, dF + \int_{M_n}^{\bar{R}} 1(\pi = 1 \mid \Delta_n, C) R_n \, dF \\ &= \int_{\underline{R}}^{M_n - \underline{\Delta}_n} R_n \, dF + \int_{M_n + \bar{\Delta}_n}^{\bar{R}} R_n \, dF, \end{aligned}$$

where $1(\cdot)$ is the indicator function, taking the value of 1 if the expression holds and 0 otherwise. The final equality follows from the definition of $\underline{\Delta}_n$ and $\bar{\Delta}_n$ and the fact that $1(\pi = 1 \mid \Delta_n, C) = 0$ when $R_n \in [M_n - \underline{\Delta}_n, M_n + \bar{\Delta}_n]$. The question is, how does $E[R_n]$ change as a function of M_n? We can differentiate this expression with respect to M_n to yield the following:

$$\frac{\partial E[R_n]}{\partial M_n} = (M_n - \underline{\Delta}_n) f(M_n - \underline{\Delta}_n) - (M_n + \bar{\Delta}_n) f(M_n + \bar{\Delta}_n).$$

Thus, we conclude that

$$\frac{\partial E[R_n]}{\partial M_n} > 0 \iff \frac{(M_n - \underline{\Delta}_n)}{(M_n + \bar{\Delta}_n)} > \frac{f(M_n + \bar{\Delta}_n)}{f(M_n - \underline{\Delta}_n)},$$

where all terms are positive. It is clear that, depending on the specific distributions chosen, this could yield either positive or negative values for this derivative.

Now, we consider the context described by Wu and Huberman (2008). This is best captured in the question they ask: "What is the point of leaving another 5-star review after one hundred people have already done so?" (p. 337). Simply put, positive reviews are not likely to be posted anymore. This implies that $(M_n + \bar{\Delta}_n)$ is very large, perhaps exceeding $\bar{R}$. Since $f = 0 \; \forall R > \bar{R}$, we know that $(M_n - \underline{\Delta}_n)/(M_n + \bar{\Delta}_n) > f(M_n + \bar{\Delta}_n)/f(M_n - \underline{\Delta}_n) = 0$, which proves condition (a).

Another context in which we can derive a similar result is when the prevailing mean is more in the center of the ratings distribution, and the indifference deviations are symmetric: $\bar{\Delta}_n = -\underline{\Delta}_n \equiv \Delta^*$. In this case, since $(M_n - \underline{\Delta}_n)/(M_n + \bar{\Delta}_n) = 1$, a sufficient condition for the positive impact of an increase in mean would be a sufficiently left-skewed ratings distribution, implying that $f(M_n + \Delta_n^*) > f(M_n - \Delta_n^*)$. A J-shaped distribution would be an example of such a distribution. This proves condition (b). □

References

Anderson, E. W. 1998. Customer satisfaction and word of mouth. *J. Service Res.* **1**(1) 5–17.

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

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

Burt, R. S. 1992. *Structural Holes: The Social Structure of Competition*. Harvard University Press, Cambridge, MA.

Çelen, B., S. Kariv. 2004. Distinguishing informational cascades from herd behavior in the laboratory. *Amer. Econom. Rev.* **94**(3) 484–497.

Çelen, B., S. Kariv. 2005. An experimental test of observational learning under imperfect information. *Econom. Theory* **26**(3) 677–699.

Chevalier, J., A. Goolsbee. 2003. Measuring prices and price competition online: Amazon.com and BarnesandNoble.com. *Quant. Marketing Econom.* **1**(2) 203–222.

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

Chintagunta, P. K., S. Gopinath, S. Venkataraman. 2010. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets. *Marketing Sci.* **29**(5) 944–957.

Dellarocas, C., R. Narayanan. 2006. A statistical measure of a population's propensity to engage in post-purchase online word-of-mouth. *Statist. Sci.* **21**(2) 277–285.

Dellarocas, C., X. Zhang, N. F. Awad. 2007. Exploring the value of online product reviews in forecasting sales: The case of motion pictures. *J. Interactive Marketing* **21**(4) 23–45.

Duan, W., B. Gu, A. B. Whinston. 2008a. Do online reviews matter?—An empirical investigation of panel data. *Decision Support Systems* **45**(4) 1007–1016.

Duan, W., B. Gu, A. B. Whinston. 2008b. The dynamics of online word-of-mouth and product sales—An empirical investigation of the movie industry. *J. Retailing* **84**(2) 233–242.

Duan, W., B. Gu, A. B. Whinston. 2009. Informational cascades and software adoption on the Internet: An empirical investigation. *MIS Quart.* **33**(1) 23–48.

Duflo, E., E. Saez. 2003. The role of information and social interactions in retirement plan decisions: Evidence from a randomized experiment. *Quart. J. Econom.* **118**(3) 815–842.

Forman, C., A. Ghose, B. Wiesenfeld. 2008. Examining the relationship between reviews and sales: The role of reviewer identity disclosure in electronic markets. *Inform. Systems Res.* **19**(3) 291–313.

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

Godes, D., D. Mayzlin. 2009. Firm-created word-of-mouth communication: Evidence from a field test. *Marketing Sci.* **28**(4) 721–739.

Granovetter, M. S. 1973. The strength of weak ties. *Amer. J. Sociol.* **78**(6) 1360–1380.

Hu, N., L. Liu, J. J. Zhang. 2008. Do online reviews affect product sales? The role of reviewer characteristics and temporal effects. *Inform. Tech. Management* **9**(3) 201–214.

Hu, N., P. Pavlou, J. Zhang. 2009. Identifying and overcoming self-selection biases in online product reviews. Working paper, Temple University, Philadelphia.

Iyengar, R., C. Van den Bulte, T. W. Valente. 2011. Opinion leadership and social contagion in new product diffusion. *Marketing Sci.* **30**(2) 195–212.

Jacoby, J., D. E. Speller, C. K. Berning. 1974a. Brand choice behavior as a function of information load: Replication and extension. *J. Consumer Res.* **1**(1) 33–42.

Jacoby, J., D. Speller, C. A. Kohn. 1974b. Brand choice behavior as a function of information load. *J. Marketing Res.* **11**(1) 63–69.

Karshenas, M., P. L. Stoneman. 1993. Rank, stock, order, and epidemic effects in the diffusion of new process technologies: An empirical model. *RAND J. Econom.* **24**(4) 503–528.

Kee, T. 2008. Majority of online shoppers check at least four reviews before buying. *Online Media Daily* (February 19), http://www.mediapost.com/publications/?fa=Articles.printFriendly&art_aid=76727.

Kumar, N., I. Benbasat. 2006. Research note: The influence of recommendations and consumer reviews on evaluations of websites. *Inform. Systems Res.* **17**(4) 425–439.

Li, X., L. M. Hitt. 2008. Self-selection and information role of online product reviews. *Inform. Systems Res.* **19**(4) 456–474.

Liu, Y. 2006. Word of mouth for movies: Its dynamics and impact on box office revenue. *J. Marketing* **70**(July) 74–89.

Lorenz, J. 2009. Universality in movie ratings distributions. *Eur. Phys. J. B* **71**(2) 251–258.

Manchanda, P., Y. Xie, N. Youn. 2008. The role of targeted communication and contagion in new product adoption. *Marketing Sci.* **27**(6) 961–976.

Manski, C. F. 1993. Identification of endogenous social effects: The reflection problem. *Rev. Econom. Stud.* **60**(3) 531–542.

McGran, K. 2005. Internet hurting dealer profits; savvy shoppers hunt for bargains buyers often know more than sellers. *Toronto Star* (November 12) G4.

Moe, W. W., D. A. Schweidel. 2012. Online product opinions: Incidence, evaluation, and evolution. *Marketing Sci.* **31**(3) 372–386.

Moe, W. W., M. Trusov. 2011. Measuring the value of social dynamics in online product ratings forums. *J. Marketing Res.* **48**(3) 444–456.

Moon, S., P. K. Bergey, D. Iacobucci. 2010. Dynamic effects among movie ratings, movie revenues, and viewer satisfaction. *J. Marketing* **74**(1) 108–121.

Moul, C. C. 2007. Measuring word of mouth's impact on theatrical movie admissions. *J. Econom. Management Strategy* **16**(4) 859–892.

Putsis, W. P., Jr., S. Balasubramanian, E. W. Kaplan, S. K. Sen. 1997. Mixing behavior in cross-country diffusion. *Marketing Sci.* **16**(4) 354–369.

Qu, Z., H. Zhang, H. Li. 2008. Determinants of online merchant ratings: Content analysis of consumer comments about Yahoo merchants. *Decision Support Systems* **46**(1) 440–449.

Rogers, E. M. 1993. *Diffusion of Innovations*, 4th ed. Free Press, New York.

Salganik, M. J., P. S. Dodds, D. J. Watts. 2006. Experimental study of inequality and unpredictability in an artificial cultural market. *Science* **311**(5762) 854–856.

Sun, M. 2010. How does variance in ratings matter? Working paper, Stanford University, Stanford, CA.

Trusov, M., R. E. Bucklin, K. Pauwels. 2009. Effects of word-of-mouth versus traditional marketing: Findings from an Internet social networking site. *J. Marketing* **73**(5) 90–102.

Van Den Bulte, C., Y. V. Joshi. 2007. New product diffusion with influentials and imitators. *Marketing Sci.* **26**(3) 400–421.

Villanueva, J., S. Yoo, D. M. Hanssens. 2009. The impact of marketing induced versus word-of-mouth customer acquisition on customer equity growth. *J. Marketing Res.* **45**(1) 48–59.

Wojnicki, A., D. Godes. 2010. Signaling success: Strategically-positive word of mouth. Working paper, University of Toronto, Toronto.

Wu, F., B. Huberman. 2008. How public opinion forms. C. Papadimitrou, S. Zhang, eds. *Internet and Network Economics*. Lecture Notes in Computer Science, Vol. 5385. Springer, Berlin, 334–341.

Zhang, J. 2010. The sound of silence: Observational learning in the U.S. kidney market. *Marketing Sci.* **29**(2) 315–335.

Zhang, X. M., F. Zhu. 2011. Group size and incentives to contribute: A natural experiment at Chinese Wikipedia. *Amer. Econom. Rev.* **101**(4) 1603–1620.

Zhu, F., X. M. Zhang. 2010. Impact of online consumer reviews on sales: The moderating role of product and consumer characteristics. *J. Marketing* **74**(2) 133–148.