Text Performance on Vine Stage? The Effect of Incentive on Product Review Text Quality

Dandan Qiao
National University of Singapore, qiaodd@nus.edu.sg

Huaxia Rui
Simon Business School, University of Rochester, huaxia.rui@simon.rochester.edu

Incentivized reviews have become increasingly prevalent on product review sites such as Amazon. While outright fake reviews are clearly unacceptable and should be removed from review platforms, reviews by incentivized consumers with otherwise authentic product experiences fall into a gray area. On the one hand, many critics and researchers have warned of their harm by pointing out their biased ratings. On the other hand, these reviews might complement organic reviews with review text of higher quality. The current paper studies whether incentivized reviews on Amazon are more coherent and offer richer detail. We use Amazon's platform-incentivized reviews, known as Vine reviews, for our primary sample, and use seller-incentivized reviews for checking robustness. Estimations from a two-way fixed-effect model consistently show that incentivized reviews do compensate for their reduced impartiality through better text quality, measured by discourse coherence and level of relevant detail. This finding is further supported by a randomized experiment using Amazon Mechanical Turk (MTurk). Hence, current literature findings on the poor quality of text of incentivized reviews, based on review length and lexical complexity only portray an incomplete picture of incentivized reviews. Given that numerical ratings for products via incentivized reviews are likely biased while their text content is of high quality, a natural way to embrace incentivized reviews is to keep their text content, suppress their numerical ratings, and always highlight the label of "incentivized reviews".

Key words: incentivized review; word-of-mouth; discourse coherence; aspect richness

1. Introduction

In the pre-internet era, consumers would learn about products from advertisements via mass media and from (offline) word of mouth comments shared by family and friends. While an advertisement is by definition incentivized with carefully crafted content, a word-of-mouth message is organic albeit less well-crafted. The arrival of the internet and the ensuing social media revolution overcame the limitation of offline word-of-mouth opinions by digitally connecting all consumers through various online platforms such as Yelp and Amazon. For a while, it was widely believed that consumer reviews, as the digital counterpart of offline consumer word of mouth, were organic and trustworthy. However, as a recent Wall Street Journal article[1] warned, more than a third of online reviews on major websites, including those on Amazon.com, Walmart.com, and Sephora.com, are generated by robots or people paid to write them. While outright fake reviews are clearly unacceptable and should undoubtedly be rooted out, reviews by incentivized consumers with otherwise authentic product experiences (henceforth referred to as incentivized reviews) fall into a gray area. These reviews are essentially a hybrid of advertisements and word-of-mouth opinions.

Should we reject these reviews by discouraging their generation and dissemination because of their incentivized nature? Or, do they have a role to play in the product review ecosystem, given their significant volume and potential values, especially for products with relatively few reviews? Existing literature has largely sounded the alarm about the biased numerical ratings of products related to these incentivized reviews, which is an important dimension of product reviews. The other dimension, the quality of review text, is certainly no less important, but is much less well-understood, partly because of the difficulty of measuring review text quality.

Given this industry background and literature status, we aim to address the following research question in the current paper.

Research Question: Does an incentive provision affect review text quality?

[1] https://www.wsj.com/articles/black-friday-shoppers-beware-of-fake-five-star-reviews-11574937001

We hypothesize that incentivized reviews are of higher text quality and therefore sit in the middle of the product information spectrum with advertisements and organic reviews at either end. Our basic intuitions are twofold. First, from the social psychological perspective, consumers who receive an incentive for writing a review are more likely to treat the production of the review as a job being "scrutinized" by their "employers" (i.e., the platform or the seller). Such a front-stage performance, using the language of Goffman (1956), naturally results in higher text quality than that of organic reviews which are more like a back-stage performance in the minds of non-incentivized reviewers. Second, from the economic perspective and assuming the objective of reviewers is influencing or gaining approval from other consumers, incentivized reviewers will naturally exert more effort in writing better quality reviews to compensate for their disadvantage of being perceived as less impartial due to the explicit label of incentivized reviews. These intuitions lead us to believe that with the activation of the "incentivized mode", a reviewer is more likely to improve the text quality of their review.

Current literature has documented certain evidence indicating that incentivized reviews tend to be shorter and use less-complex words. We depart from the literature by more directly measuring review text quality using coherence and aspect richness, both of which are important quality indicators but are not yet examined in the product review literature or used in the IS field. Coherence is directly related to inference during reading and aspect richness is important for product reviews. The construction of both measures is based on recent developments in computational linguistics.

Our main empirical setting is Amazon product reviews. We treat Amazon Vine reviews as incentivized reviews and use a two-way, fixed-effect model (i.e., product fixed effect and reviewer fixed effect) for identification. Estimation results consistently support our hypothesis, whether we use syntax-based coherence measures or semantic-based coherence measures, and regardless of how we measure aspect richness. Furthermore, we find similar results if we replace platform-incentivized reviews (i.e., Vine reviews) with seller-incentivized reviews. To complement the observational study, which is inevitably vulnerable to some concern regarding endogeneity, we conducted a randomized

experiment using MTurk, the results of which also suggest that incentivized reviews score higher in coherence and aspect richness.

Despite the legitimate concerns that incentivized reviews may be biased in the same way traditional advertisements are, our study suggests that they are also more coherently written and include more detail. Hence, we believe that incentivized reviews complement organic reviews and should have their place on review platforms *as long as* they are explicitly labeled.

The rest of the paper is organized as follows. In Section 2, we review the burgeoning literature on incentivized reviews. In Section 3, we develop our hypothesis based on two different mechanisms. Section 4 presents detailed descriptions of how we measure coherence and aspect richness. In Section 5, we report estimation results followed by various robustness checks and heterogeneous analyses. In Section 6, we conduct a randomized experiment using MTurk to further test the hypothesis. We conclude the paper in Section 7 by discussing the contributions and limitations of the current paper.

2. Research Background

There is a broad range of literature that study the use of financial incentives to increase the quantity and quality of user-generated content in a variety of settings. For example, Hsieh et al. (2010) found that for community-based question answering (CQA) platforms, financial incentives result in more answers but not higher answer quality. In a review community that allows for 'friend' connections, Sun et al. (2017) found that introducing monetary rewards can undermine contribution rates from community members who are more socially connected, while enticing members with no friends to contribute more reviews. On the other hand, reviews contributed by members without friends seem to become shorter and less helpful after the introduction of monetary rewards, while such an effect is absent for socially connected members. Kuang et al. (2019) found that when paid-for live service is enabled within a CQA community, users tend to engage in more voluntary answer contributions to match their revenue received in the paid-for live session. On Seeking Alpha, a crowd-sourced content service for financial markets, Chen et al. (2019) found that providing monetary incentives

can significantly increase users' content output, but hardly change content quality. Through a field experiment, Burtch et al. (2019) found that financial incentives increase the volume but decrease the novelty of content on Reddit.

The current paper is more closely related to the specific but important stream of literature on incentivized *product reviews*. As many review platforms and sellers use financial incentives such as cash rewards or free samples to elicit reviews from consumers, researchers are interested in and have been investigating how the provision of a financial incentive affects the quality of product reviews. Some theoretical models have been developed to understand the impact of incentives on review contribution. For example, Liu and Feng (2016) found that different market equilibrium outcomes can emerge in review contribution when incentives vary. Duan et al. (2019) designed a two-stage game-theoretical model to analyze how incentive provision affects review contributions and the seller's profits. Although they consider review quality in their model, it mostly captures review valence. Empirical works that investigate the impact of incentives on review contribution can be roughly classified into two streams, with one focusing on the direct impact of incentives on those incentivized review contributions, while the other focuses on the indirect impact of incentives on subsequent non-incentivized review contributions. We provide a summary of these works in Table 1. Our study is more related to the first stream, i.e., uncovering the impact of incentives on the quality of incentivized reviews.

While an objective quality measure is difficult to obtain, researchers in previous studies have generally focused on two aspects of review quality: *numerical rating* and *review text*. On numerical rating, the research attention is on whether incentive provision results in an upward bias. Current literature has largely validated this concern. For example, through a series of field experiments, Cabral and Li (2015) showed that providing buyers with rebates could solicit more positive feedback which benefits the seller at the expense of general consumers. Using a dataset from Taobao.com, Lin et al. (2019) found an upward rating trend in products that use free samples to solicit reviews, further suggesting a numerical product rating bias of incentivized reviews. Similarly, but measuring

positivity as the difference between the percentage of positive emotion words and the percentage of negative emotion words in a review text, Woolley and Sharif (2021) found through controlled experiments that incentive leads to review texts with more positive words relative to negative words.

Unlike the quality evaluation of numerical ratings, assessing the quality of review text has proven to be much harder. Since our study is more related to the research listed in the top panel of Table 1, we will focus on this stream of literature. Using review length and text readability as quality measures, Khern-am nuai et al. (2018) found that for a given product, the introduction of financial incentives significantly reduces review quality in the sense of shorter review and more readable text (i.e., requiring fewer years of education). However, such a decline in text quality is largely driven by the changing composition of reviewers due to the financial incentive policy. While existing reviewers seem to write product reviews of the same quality, new reviewers who join the review platform because of the financial incentive policy produce reviews of significantly lower quality compared to those from existing reviewers. These findings are alarming but not surprising given how the financial incentive is structured: the incentive is proportional to the number of contributed reviews each of which must meet the 50-character minimal length requirement. An important take-away from this study is that the design of incentive is crucial if a platform wishes to incentivize review contribution in a meaningful way.

Burtch et al. (2018) examined the effect of financial incentive, social norm, and the combination of these two, on review volume and review length. Through two randomized experiments, they found that financial incentive does not stimulate longer reviews but social norm does. Furthermore, through an observational study, they found that on Amazon, consumers tend to write shorter reviews as a result of a financial incentive provided by third-party sellers. An important distinction between Burtch et al. (2018) and Khern-am nuai et al. (2018) is that while the treatment is at the product level in Khern-am nuai et al. (2018), it is at the reviewer level in Burtch et al. (2018). Together, these two pioneering papers raise the alarm on the effect of providing financial incentives

on review text quality. The current paper continues this investigation, but differs from the literature in two important ways. First, we propose theoretical mechanisms to understand why reviewers might exert more effort in writing reviews in response to a financial incentive, leading to better review text quality. Although this implication seems to be at odds with current literature findings, such an effect is not entirely counter-intuitive. Indeed, if we consider incentivized reviews as a type of commercialized content, then the literature on ad-supported content suggests that content quality may increase as a result of financial incentives. For example, in a very different context, Sun and Zhu (2013) found that bloggers who participated in an ad-revenue-sharing program wrote blogs of a better quality, relative to those who did not participate. Second, we evaluate text quality with measures based on linguistic theory and derived from machine learning algorithms which allow us to go beyond word count and further enrich our understanding of reviewers' behavioral change in response to a financial incentive.

3. Hypothesis Development

By writing a review, a consumer becomes associated with the platform, the seller, and other consumers. Both the platform and sellers can provide incentives to consumers to write reviews. For ease of illustration, we refer to both as *incentive providers*, or simply *providers*, in this paper. To understand how such an incentive might affect the effort a consumer puts into their reviews, we propose two mechanisms, one based on the reviewer's consideration of the incentive provider, and the other based on the reviewer's consideration of other consumers. For any individual reviewers, the effect of an incentive provision on effort may be driven by either or both mechanisms. We analyze these two mechanisms separately in this section.

3.1. Provider-Oriented Mechanism

The intuition for the effect of an incentive provision on review effort is quite simple: when a consumer writes an incentivized review knowing that their review will likely be scrutinized by the incentive provider, they will put more effort into crafting the message because the review is essentially a "text performance" at the stage front. Such a phenomenon has been well-studied in what is known as dramaturgical theory in sociology.

According to this theory, people engage in "front-stage performance" when their behaviors are observed by some form of audience (Goffman et al. 1978). This is a theatrical metaphor that sociologists propose to interpret and model social interactions which share the basic elements of the dramaturgy: 1) the *setting* which defines the context or situation where a performance takes place; 2) the *front* which refers to the expressive equipment like appearances and manners that actors can manage to help achieve the performance goal, and 3) the *performance* which is the final dramaturgical realization delivered to the audience. In social interactions, all participants define a situation where individuals intentionally or unwittingly adjust their expressions in an attempt to foster a favorable impression, in much the same way as actors performing their roles in front of the audience.

Business researchers have conducted dramaturgical analyses in service marketing (Swartz and Iacobucci 2000). Production and consumption are simultaneous in service exchange, which characterizes a high degree of interactions between service providers and consumers. This interactive nature defines service delivery as a stage-like context, where impression management is the very essential goal of dramaturgical performance (Grove and Fisk 1992). Service providers, like restaurant waiters and airline attendants, must carefully manage their 'front' (e.g., proper demeanor and attitudes) to create a positive, solid image and assure consumer satisfaction. Evaluating service exchange using this dramaturgical perspective can help guide service provision (Grove and Fisk 1983, Solomon et al. 1985). For example, management consultants analyze their activities via the dramaturgical view to help coach candidates so that they can better serve clients (Clark and Salaman 1998). The essence of dramaturgical theory can be summarized as impression management. In the presence of audiences, individuals perform like actors (e.g., adjusting their appearances and manners) to control the impression they can make on others.

Writing a review in exchange for an incentive is essentially a service exchange, and hence may naturally trigger the front-stage behavior of the review writer. However, the performance is delivered completely through text and rating. Therefore, impression management is achieved through upward-biased ratings and higher-quality writing.

3.2. Consumer-Oriented Mechanism

Reviews are public goods that benefit all consumers. Many consumers write their reviews with the objective of helping other consumers to make informed buying decisions. Hence, we analyze how such a consideration in motivating review contribution is moderated by the provision of incentive. To do so, we propose a stylized model and assume the reviewer's objective is to maximize the approval rate of their review by other consumers. For example, the reviewer might hope to win many helpful votes from other consumers.

We start by modeling the value of a review which depends both on the impartiality of the review and the quality of the review. An impartial review makes it easy for consumers to evaluate the quality of the associated product, and hence such a review is considered more helpful. On the other hand, the quality of a review, such as its coherence and aspect richness, is also an important dimension to the value of a review. Therefore, we model the value of a review using the Cobb-Douglas production function

$$v(w,r) = w^{\theta} \cdot r^{1-\theta}, \quad \theta \in (0,1) \tag{1}$$

where the two inputs, w and r, are review quality and review impartiality, respectively, and the parameter θ represents the output elasticity of review quality w.

A consumer considers a review helpful only if the value of the review exceeds a certain threshold h. Clearly, different consumers have different thresholds. We model h as a random variable drawn from some distribution. Because power law distributions have been widely used in physical, biological, and social sciences to model the heterogeneity of magnitudes, we assume a Pareto distribution for h, with scale parameter h_m and shape parameter α. The percentage of consumers who would find the review helpful is thus $\Pr(v \geq h)$ which is the cumulative distribution function (CDF) of the Pareto distribution

$$F(v) \equiv \begin{cases} 1 - (\frac{h_m}{v})^{\alpha}, & v \geq h_m \\ 0, & v < h_m \end{cases} \tag{2}$$

Suppose the cost for writing a review of quality w is $C(w)$ which is an increasing and convex function. The reviewer's decision problem is choosing the best w so as to maximize the objective function $U(w,r) \equiv F(v(w,r)) - C(w)$. In our context, we treat the impartiality r as exogenous rather than a decision variable. Relating to our empirical setting, we can think of r as a dichotomous variable whose value is larger for an organic review than for an incentivized review.

From the first-order condition, we have $F'(u(w,r)) = C'(w)$. With the Cobb-Douglas production function, i.e., Equation (1), we have

$$F'(w^{\theta} \cdot r^{1-\theta}) r^{1-\theta} \theta w^{\theta-1} = C'(w).$$

With the Pareto distribution, i.e., Equation (2), we obtain

$$\frac{\alpha h_m^{\alpha}}{w^{\theta(\alpha+1)} r^{(1-\theta)(\alpha+1)}} r^{1-\theta} \theta w^{\theta-1} = C'(w) \Longrightarrow \frac{\theta \alpha h_m^{\alpha}}{w^{1+\alpha\theta} r^{\alpha(1-\theta)}} = C'(w)$$

Hence, the optimal writing quality w^* as a function of impartiality r is determined by

$$\theta \alpha h_m^{\alpha} = r^{\alpha(1-\theta)} C'(w^*) \left(w^*\right)^{1+\alpha\theta} \tag{3}$$

Finally, taking the derivative of (3) with respect to r, we have

$$\frac{dw^*}{dr} = -\frac{\alpha(1-\theta) w^{1+\alpha\theta} C'(w)}{r \frac{d}{dw}\left(w^{1+\alpha\theta} C'(w)\right)}\Bigg|_{w=w^*}.$$

Because $\alpha > 0$, $\theta \in (0,1)$, and C is increasing and convex, we immediately see that $dw^*/dr < 0$, hence we have the following proposition.

Proposition: The optimal text quality w^* is a decreasing function of the review impartiality r.

To summarize, whether the consideration is oriented towards the incentive provider or other consumers, the effect of an incentive provision on the text quality of a review is positive. Therefore, we propose the following hypothesis for empirical testing.

Hypothesis: Receiving a financial incentive causes a reviewer to write a higher-quality review text.

There are two challenges to the empirical test of the above hypothesis. First, how do we measure the text quality of a product review? Second, how do we identify the effect of an incentive provision

in the absence of an ideal experiment where researchers can randomly assign an incentive provision among reviewers? We devote the next section to addressing the first challenge before discussing and executing our empirical strategy in Section 5.

4. Measurements

Wang and Strong (1996) proposed a conceptual framework to assess data quality on multiple dimensions. We propose two quality measures for review texts that are consistent with this conceptual framework in terms of understandability/interpretability and completeness. The concept of discourse coherence in linguistics intends to capture the extent to which inference load is needed when we read a discourse (Grosz et al. 1995). Higher coherence means better understandability/interpretability, hence it is a good proxy for text data quality assessment. This is especially true for review texts because the evaluative nature (Carenini et al. 2013) of a review makes the smoothness of logical inference particularly relevant. Another important dimension in the assessment framework of Wang and Strong (1996) is completeness—the extent to which data is of sufficient breadth, depth, and scope for a task. The measure of aspect richness is consistent with this notion, hence is a good proxy for completeness. Indeed, we tend to associate high-quality reviews as those presenting complete and detailed comments on a product reflecting multiple aspects of the product.

We elaborate in this section on the details of measure construction for discourse coherence and aspect richness.

4.1. Discourse Coherence

As evaluative texts, it is important for a review to demonstrate logical coherence to be convincing and persuasive. In linguistics, discourse coherence refers to "*the quality of being logically integrated, consistent, and intelligible*" (Stein and Glenn 1975), and indicates how well information is connected in verbal or written discourses (Foltz et al. 1998). Coherence is a foremost quality proxy for textual discourses (Witte and Faigley 1981). A coherent discourse can reduce the inference load during the comprehension (Grosz et al. 1995). Many educational institutes evaluate student essays at least partially based on coherence (Higgins et al. 2004).

The most influential framework for theorizing discourse coherence is the centering theory (Grosz et al. 1995), the basic premise of which is that the smooth shift of attentional states across utterance determines the degree of coherence of a discourse. The attentional state models the focus of attention for users in discourse construction and interpretation. This centering principle is a cornerstone for our understanding of the mechanism of discourse coherence, guiding many studies after its establishment (Poesio et al. 2004).

Researchers have developed different methods to measure discourse coherence, which can be roughly categorized into the syntactic approach and the semantic approach. With the syntactic approach, coherence is modeled as how entities are introduced and discussed in discourses (Karamanis et al. 2004, Elsner and Charniak 2011). Consistent with the centering principle, these entities indicate the attentional centers, while choices of referring expressions could help sustain the connectedness. Earlier attempts to implement this approach relied on predefined knowledge and manually annotated corpus (Reiter and Dale 2000), which suffer from the scalability issue. Barzilay and Lapata (2008) proposed an automatic approach called the entity-grid method. This method derives an entity-grid representation for each discourse, where columns correspond to entities with different saliences, and rows correspond to sentence sequences within the investigated discourse. Across-utterance center transitions under the centering theory are therefore projected to entity transfer at the sentence level. Such an entity-grid representation enables automatic learning of the distributed patterns of entity transition and hence can be easily used to measure coherence.

The semantic approach can be traced back to Halliday and Hasan (1976) who argue that the relatedness of lexical meaning across textual units is the fundamental property of discourse coherence. To model such semantic relatedness, Foltz et al. (1998) developed an LSA-based method to automatically vectorize the meaning of each sentence and measure the relatedness of adjacent sentences by the cosine similarity of their vector representations. More specifically, for a discourse consisting of n sentences, its semantic coherence is measured as the average similarity of each neighboring sentence pair, as formulated below,

$$\text{Semantic Coherence} = \frac{\sum_{i=1}^{n-1} \cos(\vec{S}_i, \vec{S}_{i+1})}{n-1} \tag{4}$$

where $\vec{S}_i$ and $\vec{S}_{i+1}$ denotes the vector representation of the i-th sentence and the $(i+1)$-th sentence, respectively.

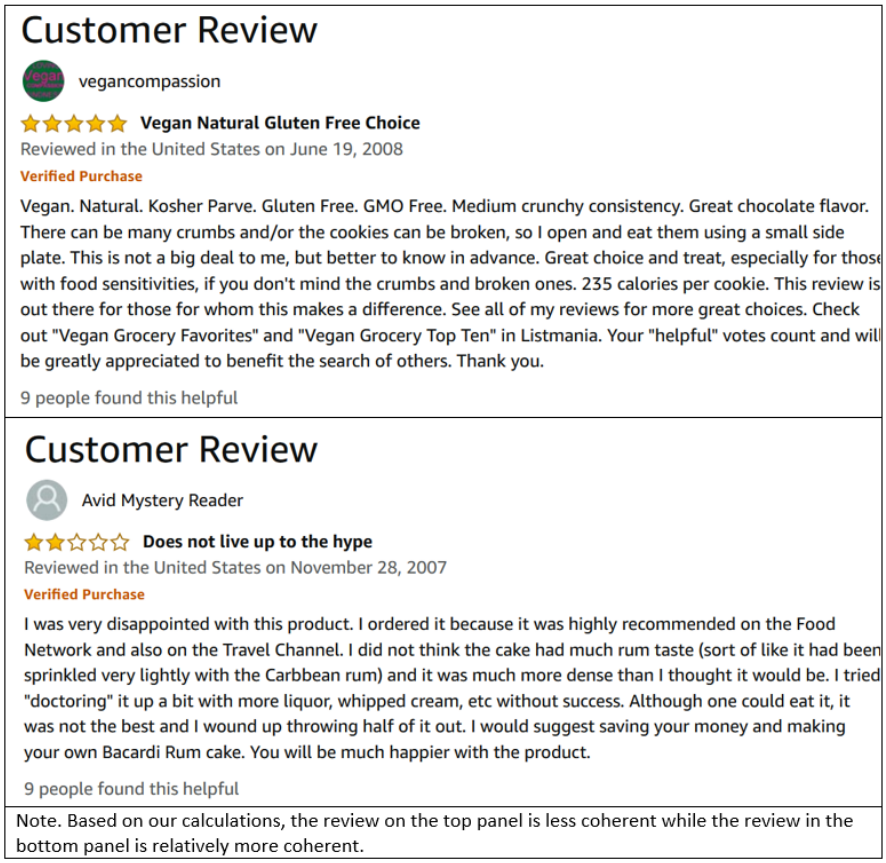
For the current study, we use both the syntactic approach and the semantic approach to measure coherence. For the syntactic approach, we use the entity-grid method. The entity-grid method relies on the model likelihood value to measure the coherence of each review text. Although it is difficult to interpret the magnitude of the syntactic coherence, such a coherence value enables us to rank reviews, similar to the ordinal principle for measuring individual utilities in economics. The ordinal approach also well suits the platform's needs, i.e., ranking reviews based on text quality to reduce inference load. Figure 1 shows two examples of reviews with different syntactic coherence scores. For the semantic approach, we replace the LSA semantics with *word2vec* which is more up to date and has been widely used in many disciplines. A sentence is represented as the average embedding of the words in the sentence, which is then used to compute the semantic coherence based on Equation (4).

4.2. Aspect Richness

Product reviews are a unique type of discourse because they are expected to convey information on multiple product attributes. Reviewers express opinions on these product attributes, which constitute the basis for other consumers to evaluate a product. In text analytics, product attributes are referred to as aspects. Whether a review covers many aspects of a product is an important indicator of the review text quality.

The main challenge of measuring aspect richness is the extraction of aspect terms. The literature has proposed three general approaches for this task: rule-based, supervised, and unsupervised. The rule-based approach identifies aspect terms using a combination of frequent nouns and noun phrases, dependency parsing, and opinion lexicon. This approach works well only when the aspect terms are restricted to a small set of nouns. The supervised approach relies on manual annotation and suffers from the domain adaptation issue. Because our data consists of a large number of reviews on a variety of different product types, an unsupervised approach is more promising.

Figure 1 Examples of Less Coherent Review (Top) and More Coherent Review (Bottom)



Note. Based on our calculations, the review on the top panel is less coherent while the review in the bottom panel is relatively more coherent.

Most algorithms for unsupervised aspect extraction are based on variants and extensions of Latent Dirichlet Allocation (LDA). However, conventional LDA models encode word co-occurrence information only at the document level and are less efficient at identifying aspect terms. He et al. (2017) proposed the attention-based aspect extraction (ABAE) model to overcome the limitation of LDA-based aspect extraction algorithms. The objective of the ABAE model is to learn a set of aspect embeddings, where each aspect can be interpreted by examining its nearest words (representative words) in the embedding space. More specifically, the ABAE model operates at the sentence level, by first constructing a sentence embedding vector $z_s = \sum_{i=1}^{n} a_i e_{w_i}$ where e_{w_i} is the word embedding vector for the i-th word in the sentence and the weight a_i is computed by an attention model. Second, the sentence-embedding vector is reconstructed using aspect embedding $r_s = T' p_t$ where T is the aspect-embedding matrix representing the embedding vector for each aspect, and p_t is the weight vector representing the probability that the input sentence belongs to the related

aspect. The neural network is trained to minimize the reconstruction error. By combining attention mechanisms and neural networks, He et al. (2017) demonstrated the superior performance of an ABAE model compared to other aspect extraction methods. Following the success of the ABAE model, researchers have found many applications for tasks such as recommendation justification (Ni et al. 2019), review summarization (Angelidis and Lapata 2018), and fine-grained sentiment analysis (Wang et al. 2018).

We use the ABAE model to derive a measure for aspect richness. Specifically, we derive two variables from the output of the ABAE model. The first one counts the number of aspect terms each review contains. For each aspect, the ABAE model generates a set of aspect terms that are close to the aspect embedding in the embedding space. These terms can be interpreted as the linguistic embodiments of the abstract concept of an aspect. The more aspect terms a review contains, the more detailed that aspect is covered by the review. Our second variable is based on the number of aspects a review touches on where we use a threshold to determine whether an aspect is sufficiently covered in the review.[2] In the appendix, we provide examples of aspect terms in the category "Electronics". For ease of understanding, we name each aspect through manual inspection of the corresponding aspect terms. The example is based on ten aspects with 50 terms in each aspect.

4.3. Measure Validation

Although both measures have been validated in the literature in other contexts, we further check the validity of the proposed metrics in our specific context. To do so, we randomly selected 1,000 pairs of reviews, each of a similar length and within the same category. For each pair of reviews, we recruited three annotators to answer the questions: "Which review is more coherent?" and "Which review is more detailed?". After aggregating such annotations for each pair of reviews, we compare the coherence annotation with our own computation using the kappa statistic, which is a popular method to measure inter-rater reliability (McHugh 2012). The kappa statistic is 0.87 for coherence

[2] The ABAE model maps each review to a distribution over all aspects. When a review's probability on one particular aspect is larger than the median value of all probability values, we consider the aspect covered by the review.

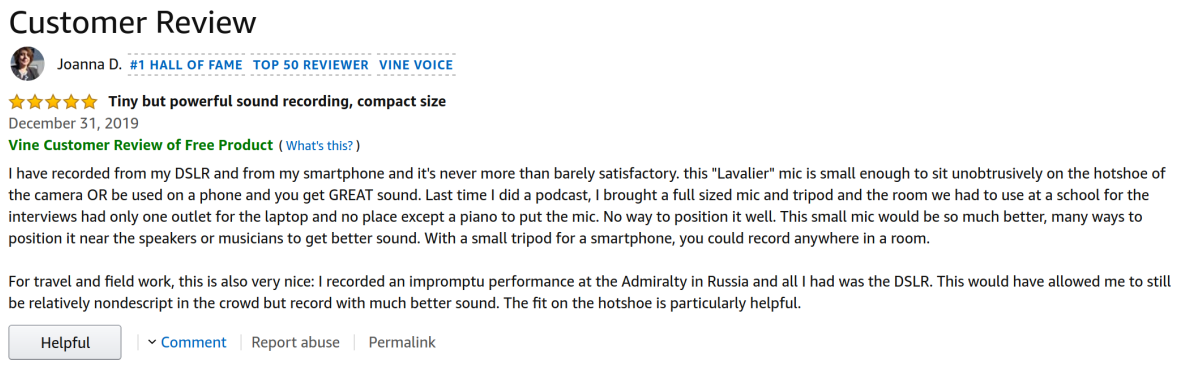
and is 0.85 for detailedness. Since a value over 0.8 indicates a strong agreement between raters (human annotation and our computation), we believe the proposed metrics capture coherence and aspect richness well for review texts.

5. Observational Study

We use product reviews on Amazon for the observational study. The full dataset includes all reviews available on Amazon from the beginning until August 2015. We focus on Vine reviews[3] because these constitute the most important type of incentivized review on Amazon and their clear labeling by Amazon also ensures that our analyses are less susceptible to measurement errors. In one of our robustness checks, we also used text mining to construct an alternative sample by detecting third-party incentivized reviews to further test the main findings of our study.

To participate in Amazon's Vine program, sellers first register their products in the program. Amazon then distributes free samples of these products to consumers who will post reviews after receiving and assessing the products. Direct contact between sellers and reviewers is prohibited to minimize the possibility of collusion. Figure 2 shows an example of a Vine review. Because consumers write Vine reviews in exchange for products given to them for free, these Vine reviews are clearly incentivized reviews.

Figure 2 An Example of Vine Review



[3] https://www.amazon.com/gp/vine/help

For an ideal experiment for our empirical question, we needed to randomly assign some consumers to the treatment group who write product reviews in exchange for some financial incentive, and other consumers to the control group who would then write organic reviews without a financial incentive. Such a field experiment is difficult to conduct in practice. Hence, we resorted to a two-way fixed-effect model for identification. We essentially exploited the fact that many reviewers write multiple reviews and many products receive multiple reviews. Therefore, including both reviewer fixed effects and product fixed effects in a panel data model can alleviate endogeneity concerns due to unobserved reviewer characteristics or unobserved product characteristics. Because of this identification strategy, only reviews whose authors have written at least one organic review and one Vine review can contribute to the coefficient estimation for our main variable. Similarly, only reviews for which the associated products have received at least one organic review and one Vine review will contribute to the estimation. After applying these criteria to the full dataset, we obtained our main sample. Table 2 reports summary statistics for the data at review level, reviewer level, and product level. Table 3 reports summary statistics for Vine reviews in particular. Table 4 reports the correlations between the proposed metrics and other review-relevant variables.

5.1. Main Results

The two-way fixed effect model can be formulated as

$$Y_{ij} = \beta Incentive_{ij} + \theta_i + \phi_j + X_{ij} + \varepsilon_{ij} \tag{5}$$

where i, j indexes product and reviewer respectively.

The dependent variable Y_{ij} denotes review text quality, either measured by coherence or aspect richness. Our main variable of interest is the binary variable $Incentive_{ij}$ indicating whether a review is a Vine review. The product fixed effect is captured by θ_i and the reviewer fixed effect is captured by ϕ_j. The vector of control variables X_{ij} includes review length, year fixed effects, and regional market fixed effects. Review length is an important control variable for the measurement of coherence and aspect richness, and it has been used in the literature as an alternative indicator

of review text quality. We are interested in the estimate of β which captures the effect of the incentive provision on coherence or aspect richness. We clustered standard errors both by product and by reviewer, so we allowed for error correlation both across reviews of the same product and across reviews by the same reviewer.

Table 5 reports the estimation results. Column (1) reports the estimation results when the dependent variable is syntactic coherence and Column (2) reports the estimation results when the dependent variable is semantic coherence. For both coherence measures, we found significant and positive coefficients of $Incentive$, suggesting that incentive provision leads to higher coherence level of the review text.

Columns (3) through (6) report estimation results when the dependent variable is aspect richness. Whether we used ten or twenty aspects and whether we measured aspect richness by the number of aspect terms or the number of covered aspects, the results consistently indicated that reviewers respond to incentive by providing a more detailed product description in the review text. For example, the estimated coefficient of $Incentive$ in Column (3) suggests an increase of $\exp(0.0768) - 1 \approx 7.98\%$ in terms of the number of words about product attributes due to incentive provision.

In summary, we find strong support for our hypothesis based on estimation results from the two-way fixed-effect model using Amazon's Vine review data.

5.2. Robustness Check: Time-Varying Factors

The main threat to the two-way fixed-effect identification strategy is the potential correlation between incentive provision and some time-varying confounding factors. For example, if reviewers are inclined to write more-coherent reviews over time and also are more likely to write Vine reviews over time, then the estimated coefficient of $Incentive$ would be upward biased for the coherence regression. On the other hand, if a product is more likely to receive Vine reviews during a certain period of time and consumers are less likely to write detailed reviews during such a period, then the estimated coefficient of $Incentive$ would be downward biased in the regression for aspect richness.

To alleviate these concerns, we introduce two additional control variables, $UserExperience_{ij}$ and $ProductExperience_{ij}$, in the regressions. The variable $UserExperience_{ij}$ is calculated as the

total number of reviews that reviewer j has previously written before posting a review of product i. The variable $ProductExperience_{ij}$ is calculated as the total number of reviews product i has received before it is reviewed by reviewer j.

Table 6 reports the estimation results after we include these two control variables. The positive and statistically significant estimates for the coefficient of $Incentive$ suggest that our main findings are robust.

5.3. Robustness Check: Propensity Score Matching

Matching is often used to reduce model sensitivity of regression analysis resulting from extrapolation over data ranges that do not include both groups. By balancing the distributions of observed covariates across the treated and control groups, matching may also potentially improve the balance of unobserved covariates. Because we have included both reviewer and product fixed effects in our analysis, matching can potentially alleviate concerns for unobserved time-varying confounding factors that differ across Vine reviewers and non-Vine reviewers. So, the goal of matching in our study is to ensure Vine reviewers and non-Vine reviewers are comparable.

Propensity Score Matching (PSM) is a popular technique that matches treated units (i.e., Vine reviewers in our study) to control units (i.e., non-Vine reviewers) based on the estimated propensity score, that is, the probability of becoming a Vine reviewer. To improve covariate balance at the reviewer level using PSM, we use the average characteristics of a Vine reviewer's reviews before their first Vine review to do the matching. Because the treatment and control groups should share a common support in terms of the propensity score, we discarded observations that lie outside of the common support region based on the Minima and Maxima comparison (Caliendo and Kopeinig 2008). We set the caliper value to 0.2 and adopted nearest neighbor matching with replacement to find one comparable control unit for each treated unit. Out of 7,938 Vine reviewers, we are able to match 6,761 successfully. We report in Table 7 the balance test results before and after the matching. As we can see, there is a significant improvement in covariate balance between Vine reviewers and non-Vine reviewers after the matching. Using the matched sample, we re-ran

the two-way fixed effect model. The estimations are reported in Table 8. Again, we qualitatively obtained the same findings that offering an incentive leads to a significant improvement of review textual quality, measured by coherence and aspect richness.

5.4. Robustness Check: Entropy Balance Matching

While PSM is a classical and popular matching technique, several new matching methods have been proposed in recent years among which the Entropy Balancing (EB) technique (Hainmueller 2012) is particularly effective in terms of achieving covariate balance. The EB technique relies on a maximum entropy reweighing scheme to produce a more balanced sample. As another robustness check and to reduce the reliance on PSM, we re-estimated the two-way fixed-effect model using the weighed sample generated by EB. The set of matching covariates is the same as that used in PSM. Table 9 reports the balance test results where we find an even more significant improvement of covariate balance. Table 10 reports the estimation results. Again, the results remain qualitatively the same.

5.5. Robustness Check: Verified Purchase Subsample

While Vine reviews are straightforward to identify thanks to the label provided by Amazon, there is no label for organic reviews. In our main analysis, we treat all non-Vine reviews as organic reviews. Such a labeling method is not perfect because some non-Vine reviews might have resulted from incentives provided by sellers. However, the "contamination" of organic reviews by incentivized reviews may only result in an underestimation of the causal effect we are trying to estimate. Nevertheless, we explore this issue by restricting our sample to reviews from verified purchases[4], along with those Vine reviews. This sample restriction reduces the sample size by almost a half, but can provide an insight into the full measure of the effect of an incentive provision on review text quality.

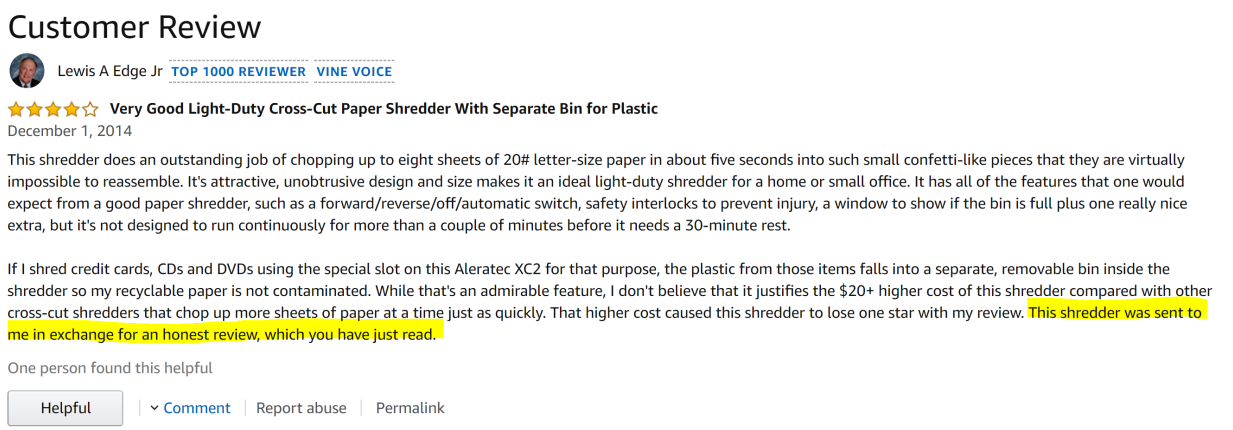
[4] According to Amazon, an "Amazon Verified Purchase" review means Amazon has verified that the person writing the review purchased the product on Amazon and did not receive the product at a deep discount. See https://www.amazon.com/gp/help/customer/display.html?nodeId=202076110.

Table 11 reports the estimation results. Again, we find, qualitatively, the same results. More interestingly, we find that all coefficient estimates are much larger in magnitude than those reported in Table 5, which confirms our intuition and suggests that our estimation based on the full sample in the main analysis is likely conservative.

5.6. Robustness Check: Third-Party Incentive

As we previously discussed, there are two types of incentivized reviews, those incentivized by the platform and those incentivized by sellers. Thus far, we have used platform-incentivized reviews (i.e., Amazon Vine reviews) to test our hypothesis. Before October 2016, Amazon also allowed seller-incentivized reviews as long as a reviewer discloses in the review text his or her relationship with the seller. Figure 3 shows an example of such a review where the reviewer received a free sample in exchange for the writing of a review. In this robustness check, we replace Vine reviews with seller-incentivized reviews to check the robustness of our main finding with this different structure of incentive provision.

Figure 3 An Example of Third-Party Incentivized Review



The key challenge of this test was to identify reviews incentivized by sellers. Unlike Vine reviews which are clearly labeled by Amazon, we had to label seller-incentivized reviews by analyzing the review text. To do so, we first randomly selected 10,000 non-Vine reviews. We then hired 15 annotators to manually extract self-disclosures indicating seller-provided incentives. Based on these

disclosures, we designed regular expressions that could be used to automatically detect similar disclosures. Testing based on another set of randomly selected 10,000 non-Vine reviews suggests that the accuracy of this classifier is above 90%.

After removing all Vine reviews from the main sample, we labeled each remaining review as incentivized if our classifier detected any disclosure of incentive provision by the seller. Table 12 reports the two-way fixed effect estimation results based on this alternative sample. Consistent with our main analyses, we again find qualitatively the same results, thereby further supporting our hypothesis.

5.7. Heterogeneity Analysis: Review Extremity

Extreme reviews are strongly opinionated. Expectations from the platform, the seller, or other consumers, are naturally higher for extreme reviews, both in terms of coherence and aspect richness. Indeed, to make a strong case for or against a product, an incentivized review writer needs to deliver a particularly coherent argument and probably covering many details. On the other hand, extreme reviews are associated with intense emotion which may induce less coherent arguments. Based on these intuitions, we conjecture that extreme reviews are less coherent in general but such an effect is attenuated for incentivized reviews. To test this, we first defined a review as being extreme if the numerical rating was either 1 or 5 (Mudambi and Schuff 2010). We then included the corresponding binary variable *Extreme* and its interaction with *Incentive* in the regression.

Table 13 reports the estimation results. Consistent with our intuitions, the negative coefficients of *Extreme* in Column (1) and Column (2) suggest that non-Vine reviews that are extreme are less coherent. However, the positive and statistically significant coefficients for the interaction term suggest that incentive provision does make an extreme review less incoherent compared with an otherwise similar but not incentivized review. Interestingly, we find that although extreme organic reviews appear to contain fewer details than non-extreme organic reviews, such a relationship seems to be reversed for incentivized reviews.

5.8. Heterogeneity Analysis: Search Versus Experience Goods

Search goods have more objective attributes compared with experience goods. It is interesting to explore whether such a difference moderates the effect of an incentive provision. Following the previous literature (Mudambi and Schuff 2010), we categorized products into search goods (e.g., office products, electronics) and experience goods (e.g., video DVD, video games, and beauty). We included the binary variable *Search* and its interaction with *Incentive* into the regression.

Table 14 reports the estimation results. While we do not find any significance in the coefficients of the variable *Search*, we do find significantly positive coefficients of the interaction term. It seems that the effect of the incentive provision is particularly stronger for search goods. One interpretation is that there are more objective attributes of search goods for incentivized reviewers to write about, which provides more room for them to perform and to demonstrate the value of their otherwise less impartial reviews.

5.9. Additional Analysis: Review Length

While content word count has been used in the literature as a proxy for review text quality (Khern-am nuai et al. 2018), it is clearly a crude one and likely has a low signal-to-noise ratio. Nevertheless, because of its simplicity and convenience, it is currently widely used and could serve as a good benchmark for comparison with the literature. Hence, we re-estimate our econometric model with review word count as the dependent variable. Table 15 reports the results. As we can see, the results are not consistent across different samples. For the main sample, we find negative and significant coefficients for *Incentive*, suggesting that Vine reviewers write shorter reviews than non-Vine reviewers. However, if we only use reviews from verified purchases, the estimated coefficient becomes significantly positive, with more than twice the magnitude. Similarly, we find a significantly positive coefficient for *Incentive* if we use third-party incentivized reviews. These inconsistent results suggest that a simple word count may be too noisy a measure for review text quality.

6. Experimental Study

The strength of causal inference in our observational study is only as good as the underlying identification assumption. While we believe the two-way fixed-effect estimation and various robustness checks alleviate many endogeneity concerns, time-varying unobserved confounding factors may still threaten the internal validity. Inspired by many recent experimental studies using MTurk, we conducted randomized experiments using MTurk to further test our hypothesis. MTurk is a marketplace where requesters publish human intelligence tasks (HITs) and workers collect a reward for completing a task. The easy access to a large subject pool on MTurk has attracted researchers across various disciplines to conduct experiments on this platform. As per our study purpose, we published a series of HITs on MTurk, where each participant was paid \$0.3 upon the completion of one HIT. To complete the task, a participant was asked to first watch a video snippet in which a Youtuber describes the use experience of a product and then to answer some questions. After completing the abovementioned required task, each participant was invited to write a review about the product. The review writing part was disclosed to the participants as a voluntary contribution. All participants, upon accepting the HIT, were randomly assigned into either the treatment or the control group. For the treatment group, referred to as the "With-Payment" group, we paid the participants \$0.5 for writing a review. For the control group, referred to as the "No-Payment" group, there was no incentive for writing a review. Participants were unaware of the two groups and treated participants were only informed about the incentive when they were invited to write a review.

To make sure that the participants were aware of the incentive administered to the review writing process, we asked the participants to write down the money they would receive at the end of the experiment. For a participant in the treated group, the correct amount should have been \$0.8, while for a participant in the control group the correct amount should have been \$0.3. Submissions with wrong answers were considered to have failed the manipulation check and were excluded from further analyses. The step-by-step experiment procedure is reported in Figure 4. Note that the

comparison between the treated and control group is only for the review writing section, which is separate from the HIT task itself (i.e., watching the video and answering questions to receive the payment of \$0.3). To highlight the distinction, we emphasized the voluntary nature of review writing in the control group and reminded participants in the treated group that the bonus \$0.5 was specifically provided for review contribution.

To increase external validity, we conducted randomized experiments using two different products[5]. For each product, we recruited 2,000 participants, with 1,000 users randomly assigned to the "With-Payment" group and the remaining users assigned to the "No-Payment" group. We obtained, in total, 2,326 completed responses. The average answer time for participants to complete an HIT was 12 minutes. Table 16 reports the descriptive statistics for continuous variables and Table 17 reports the number of observations at each level of those categorical variables. In Table 18, we conducted a randomization check on these two groups of responses. No significant differences have been found between "With-Payment" and "No-Payment" groups across all variables. We hence conclude that subjects are comparable between these two groups and are appropriate for the subsequent statistical analysis.

Table 19 reports the estimation results using the experimental data. We adopted the methods introduced in Section 4.1 to measure both the syntactic coherence and the semantic coherence of these reviews. As the coefficients of *Incentive* in Columns (1) and (2) indicate, incentivized reviews are more coherent when compared with non-incentivized reviews. We also calculated the aspect richness of these reviews using the attention-based algorithm described in Section 4.2. Coefficients of *Incentive* in Columns (3) and (4) indicate that incentivized reviews are more detailed, compared with non-incentivized reviews, although the level of statistical significance is at the 10% level which might be due to the relatively smaller sample size of the experimental study compared to the observational study.

[5] The links for the two products are `https://www.youtube.com/embed/cjI7ctWXkT0` and `https://www.youtube.com/embed/dhBRdj2t3-s` respectively.

Overall, the results from the randomized experiment support the findings from the observational study. The context based on Youtube videos also complements the context of hands-on product experience, which supports the external validity of our findings.

One caveat of the experimental study is the potential entanglement of the HIT task and the review writing task. Because we have to pay subjects to watch the video regardless of their treatment status, some participants may perceive the two causal states as low-incentive ($0.3) and high-incentive ($0.8), rather than $0 and $0.5 which are more analogous to the incentive structure of our observational study. It is possible that the causal effect of financial incentive has a jump at $0 or is concave so that the effect of raising the incentive from $0 to $0.5 is greater than the effect of raising the incentive from $0.3 to $0.8. In such a case, our estimate from the experimental study is an underestimation of the corresponding causal effect defined for the observational study. So, our qualitative findings should remain the same.

7. Conclusions and Limitations

Despite criticisms of incentivized reviews being biased, this study shows that these reviews are not without merit because they are of higher text quality, at least in terms of coherence and aspect richness. The discoveries of this study provide valuable managerial implications in several aspects. First, our study offers a fresh perspective on understanding the relation between incentivized reviews and its two "extreme" counterparts (i.e., organic reviews and advertisements). Advertisements are well crafted but biased because they are specifically designed to influence consumers in favor of the advertisers, while organic reviews are mostly unbiased but are often in the form of crude and incomplete review text. Incentivized reviews are less biased than advertisements. Meanwhile, our empirical evidence suggests they are of higher textual quality than organic reviews, thereby resembling the well-crafted nature of traditional advertisements. The balance of unbiasedness and text quality suggests that as long as the incentive is properly disclosed, incentivized reviews can play a constructive role as advertising evolves in the age of social media.

Second, our study provides helpful guidelines in managing the review system. Instead of throwing incentivized reviews out of the consumer review ecosystem because of their rating bias, we

recommend retaining them but with a watchful eye. While more studies are clearly needed to understand their impact on consumer and reviewer behaviors, we believe review platforms can consider at least two remedies: (1) All incentivized reviews must be clearly and explicitly labeled so that other consumers will not confuse them with organic reviews. Amazon pioneered this with its Vine review program, but other review platforms may need to be more proactive on this front. (2) Given that numerical ratings of incentivized reviews are likely biased with no easy approach to correcting them, we recommend either not incorporating numerical ratings from incentivized reviews into the overall product rating, or simply blocking or not even soliciting numerical ratings at all from incentivized reviewers. Indeed, if the value of an incentivized review mostly comes from its review text instead of the numerical rating, it is only natural to retain the review text while discard the numerical rating. We suggest that the future commercial system to organize reviews with three components, i.e., numerical rating based solely on organic reviews, textual content from organic reviews, and textual content from incentivized reviews. To summarize our recommendation, we envision the interface of future review platforms as resembling the design shown in Figure 5.

Third, given the value of the review text, incentivized reviews can alleviate the cold-start problem for new products in the e-commerce era, thereby facilitating market competition. Launching a new product on an e-commerce platform is particularly challenging due to the lack of product reviews. Even if the new product is of high quality and it has an attractive low price, consumers may still be hesitant to make a purchase because quality is indirectly revealed by consumer reviews which are lacking for any new product. As a result, existing products equipped with large numbers of reviews may pose a significant barrier for the entry of new products, which consequently hinders market competition. Allowing clearly labeled incentivized reviews can help break down this barrier by jump-starting a new product. Once the new product accumulates a sufficient number of organic reviews to reveal its quality, the market force will take over and consumers will benefit from healthy competition among sellers.

In addition to the abovementioned managerial implications, this paper also contributes to the academic literature on incentivized reviews. First, while previous literature has compared incentivized reviews and organic reviews in terms of text length and lexical complexity, this is the first paper to examine deeper and more direct text quality measures including coherence and aspect richness. Given that these techniques from computational linguistics have not previously been employed in the IS field, their introduction in the current paper seems particularly valuable. Second, while previous literature has been inconclusive on whether incentive provision induces higher or lower text quality, our findings consistently show that the effect of an incentive provision on text quality is positive, at least when quality is measured by coherence and aspect richness. Finally, our theoretical analyses for the hypothesis development also shed new light on the mechanisms behind the effect of an incentive provision on review text quality.

There are clearly exciting future research opportunities given the findings and limitations of the current paper. First, while the two text quality measures introduced in this paper are particularly well-suited for the review text quality, future research should go beyond these measures, especially given the rapid advancement of computational linguistics and AI technologies. Second, the observational study in the current paper is only based on incentivized reviews on Amazon which is arguably the most important product review platform. Future research can examine incentivized reviews on other review platforms to further evaluate the generalizability of our findings. Third, both extrinsic and intrinsic motivations play important roles in review writing. How financial incentives affect intrinsic motivation is a particularly interesting and important direction for future research. Last, when more granular data such as product sales and costs become available, effort can be extended to understanding how incentivized reviews affect seller profits.

Table 1 Literature Review

Review Features	Volume	Rating	Length	Helpful Votes	Feedback Speed	Lexical Complexity	Video Game Features	Self-Reported Objectivity
Direct Impacts on Incentivized Reviews								
Burtch et al. (2018)		+	-					
Cabral and Li (2015)	o	+			o			
Khern-am nuai et al. (2018)		+	-	-		-	-	
Lin et al. (2019)		+						
Stephen et al. (2012)		o	o	+				o
Wang et al. (2012)				o				
Wang et al. (2016)	+	o		o				
Wang and Sanders (2019)	+	-	+					
Indirect Impacts on Non-Incentivized Reviews								
Qiao et al. (2020)	-	+	-	o		-		
Yu et al. (Forthcoming)	+	o	+	+				

Note. (1) "+" indicates significantly positive impacts. "-" indicates significantly negative impacts. "o" indicates no significant impacts. A blank cell means no analyses carried out on the corresponding variable in the particular paper. (2) The table mainly summarizes the impacts of incentives on review contribution, while not including findings about the combinative use of incentives and other interventions.

Table 2 Descriptive Statistics

	Count	Mean	Min	Max	Std.
Measures at Review Level					
Syntactic Coherence	4234594	-0.0141	-113.6842	0.6500	1.0100
Semantic Coherence	4234594	0.5097	-0.0737	1.0000	0.1107
Review Length	4234594	197.7548	1.0000	8717.0000	246.6459
Detailedness I (With 10 Aspects)	4234594	2.4706	0.0000	7.3018	0.9879
Detailedness II (With 10 Aspects)	4234594	1.8263	0.6931	2.3026	0.2127
Detailedness I (With 20 Aspects)	4234594	2.8461	0.0000	7.6158	0.9470
Detailedness II (With 20 Aspects)	4234594	2.4542	1.3863	2.8332	0.1795
Measures at Product Level					
Syntactic Coherence	899987	-0.2734	-61.9884	0.6500	1.1347
Semantic Coherence	899987	0.5276	0.0139	0.9800	0.0837
Review Length	899987	261.5039	4.0000	7973.0000	257.4327
Detailedness I (With 10 Aspects)	899987	2.7780	0.0000	7.3018	0.8406
Detailedness II (With 10 Aspects)	899987	1.8710	0.6931	2.3026	0.1896
Detailedness I (With 20 Aspects)	899987	3.1580	0.0000	7.6158	0.7930
Detailedness II (With 20 Aspects)	899987	2.5028	1.3863	2.8332	0.1694
Review Number	899987	4.7052	1.0000	8263.0000	41.2990
Vine Review Number	899987	0.7228	0.0000	272.0000	4.1513
Measures at Reviewer Level					
Syntactic Coherence	1451311	0.3783	-58.9852	0.6500	0.4656

Table 2 continued from previous page

	Count	Mean	Min	Max	Std.
Semantic Coherence	1451311	0.4752	-0.0650	1.0000	0.1240
Review Length	1451311	90.8946	1.0000	7283.0000	124.6621
Detailedness I (With 10 Aspects)	1451311	1.9085	0.0000	6.5381	0.8308
Detailedness II (With 10 Aspects)	1451311	1.7492	0.6931	2.3026	0.1961
Detailedness I (With 20 Aspects)	1451311	2.2910	0.0000	6.8865	0.8054
Detailedness II (With 20 Aspects)	1451311	2.3869	1.3863	2.8332	0.1628
Review Number	1451311	2.9178	1.0000	30721.0000	47.6980
Vine Review Number	1451311	0.4482	0.0000	3957.0000	10.6816

Table 3 Descriptive Statistics of Vine Reviews

Statistics of Vine Reviews							
Number of Vine Reviews	650,488						
Number of Reviewers with Vine Reviews	9,256						
Number of Products with Vine Reviews	33,087						
Vine Proportion	15.36%						
Vine Proportion in Search Goods	14.68%						
Vine Proportion in Experience Goods	15.74%						
Statistics of Vine Proportion	Count	Mean	Min	Max	Std.		
Vine Proportion Per Reviewer	8,077	0.3199	0.0007	0.9859	0.2303		
Vine Proportion Per Reviewer-Category	55,290	0.4648	0.0003	0.9968	0.2511		
Vine Proportion Per Product	28,313	0.4991	0.0024	0.9714	0.2778		
Statistics of Sentiment/Rating	Positive	Negative	1	2	3	4	5
Vine Reviews	89.54%	10.46%	1.78%	5.39%	16.16%	36.38%	40.28%
Non-Vine Reviews	85.94%	14.06%	3.24%	4.60%	11.20%	27.93%	53.02%

Note.

(1) The Vine proportion is calculated as the percentage of Vine reviews versus all reviews in the corresponding classification. For example, "Vine Proportion Per Reviewer" is the average of Vine proportion over all reviewers who have written both Vine and non-Vine reviews. The calculations are similar for other classifications in this panel.

(2) "Sentiment" is calculated based on reviews' textual content using the Python package "VADER", which is a popular method for sentiment measurement.The computation of sentiments and ratings is based on reviews from reviewers who have written both Vine and organic reviews.

Table 4 Correlations Between Variables

	Syntactic Coherence	Semantic Coherence	Detailedness-I	Detailedness-II	Review Length	Helpfulness	Star Rating
Syntactic Coherence	1						
Semantic Coherence	-0.0693	1					
Detailedness-I	-0.6746	0.2169	1				
Detailedness-II	-0.1647	0.0232	0.2829	1			
Review Length	-0.7539	0.2542	0.8745	0.2397	1		
Helpfulness	-0.1078	0.0312	0.1482	0.0417	0.1685	1	
Star Rating	-0.0439	0.0047	0.0526	0.0389	0.0459	0.1741	1

Note. The measurement of helpfulness might not be very accurate, since reviews are known to have the scarcity problem of receiving votes.

Table 5 Main Analysis: Incentive Effect on Writing Quality

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0147***	0.0111***	0.0768***	0.0055***	0.0910***	0.0056***
	(0.0021)	(0.0005)	(0.0044)	(0.0008)	(0.0045)	(0.0008)
Observations	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808
Adjusted R-squared	0.9571	0.3188	0.7919	0.5956	0.7985	0.7233
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Note. For each aspect, we selected top 100 words in the embedding vector to measure Detailedness.We also vary the number of words used in the embeddings to 50, 70, and 90, which give us consistent results.

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 6 Robustness Check: Accounting for User/Product Experience

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0140***	0.0111***	0.0765***	0.0055***	0.0913***	0.0059***
	(0.0020)	(0.0005)	(0.0044)	(0.0008)	(0.0044)	(0.0007)
Observations	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808
Adjusted R-squared	0.9571	0.3189	0.7920	0.5956	0.7986	0.7234
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES
User Experience	YES	YES	YES	YES	YES	YES
Product Experience	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 7 Propensity Score Matching: Balance Test Between Treated (Vine Reviewers) and Control (Non-Vine Reviewers) Groups

	Pre-Matching			Post-Matching		
	Mean Control	Mean Treated	Standardize Difference	Mean Control	Mean Treated	Standardize Difference
Average Log Helpful Votes	0.4516	1.6260	1.8240	1.5298	1.5421	0.0198
Average Rating	4.0069	4.1622	0.2913	4.1625	4.1742	0.0214
Average Syntactic Coherence	0.3866	-0.1745	-0.7069	-0.1037	-0.1214	0.0258
Average Semantic Coherence	0.4709	0.5365	0.9610	0.5315	0.5320	0.0083
Average Log Length	3.9805	5.0752	1.6672	5.0485	5.0467	0.0029
Average Detailedness I (With 10 Aspects)	1.8647	2.7375	1.4181	2.7362	2.7179	0.0298
Average Detailedness II (With 10 Aspects)	1.7046	1.8724	1.4008	1.8682	1.8720	0.0301
Average Detailedness I (With 20 Aspects)	2.2441	3.0523	1.3368	3.0522	3.0342	0.0302
Average Detailedness II (With 20 Aspects)	2.3147	2.4829	1.3707	2.4761	2.4799	0.0291

Table 8 Robustness Check: Using Propensity Score Matching

	(1)	(2)	(3)	(4)	(5)	(6)
			With 10 Aspects		With 20 Aspects	
VARIABLES	Syntactic Coherence	Semantic Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0158***	0.0088***	0.0689***	0.0068***	0.0770***	0.0057***
	(0.0024)	(0.0006)	(0.0049)	(0.0010)	(0.0048)	(0.0010)
Observations	1,122,082	1,122,082	1,122,082	1,122,082	1,122,082	1,122,082
Adjusted R-squared	0.9624	0.2323	0.7816	0.6174	0.7847	0.7404
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 9 Entropy Balance Matching: Balance Test Between Treated (Vine Reviewers) and Control (Non-Vine Reviewers) Groups

	Pre-Matching			Post-Matching		
	Mean Control	Mean Treated	Standardize Difference	Mean Control	Mean Treated	Standardize Difference
Average Log Helpful Votes	0.4516	1.6260	1.8240	1.6260	1.6260	0.0000
Average Rating	4.0069	4.1622	0.2913	4.1622	4.1622	-0.0001
Average Syntactic Coherence	0.3866	-0.1745	-0.7069	-0.1752	-0.1745	0.0009
Average Semantic Coherence	0.4709	0.5365	0.9610	0.5364	0.5365	0.0003
Average Log Length	3.9805	5.0752	1.6672	5.0750	5.0752	0.0002
Average Detailedness I (With 10 Aspects)	1.8647	2.7375	1.4181	2.7375	2.7375	0.0001
Average Detailedness II (With 10 Aspects)	1.7046	1.8724	1.4008	1.8724	1.8724	0.0003
Average Detailedness I (With 20 Aspects)	2.2441	3.0523	1.3368	3.0522	3.0523	0.0001
Average Detailedness II (With 20 Aspects)	2.3147	2.4829	1.3707	2.4829	2.4829	0.0002

Table 10 Robustness Check: Using Entropy Balance Matching

	(1)	(2)	(3)	(4)	(5)	(6)
			With 10 Aspects		With 20 Aspects	
VARIABLES	Syntactic Coherence	Semantic Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0142***	0.0086***	0.0717***	0.0071***	0.0803***	0.0061***
	(0.0022)	(0.0005)	(0.0044)	(0.0009)	(0.0044)	(0.0008)
Observations	2,255,791	2,255,791	2,255,791	2,255,791	2,255,791	2,255,791
Adjusted R-squared	0.9558	0.1859	0.7582	0.5813	0.7616	0.7168
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 11 Robustness Check: Using Verified-Purchase Subsample

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0252***	0.0132***	0.1427***	0.0100***	0.1659***	0.0122***
	(0.0028)	(0.0007)	(0.0060)	(0.0013)	(0.0059)	(0.0012)
Observations	1,323,335	1,323,335	1,323,335	1,323,335	1,323,335	1,323,335
Adjusted R-squared	0.9534	0.3381	0.7803	0.5413	0.7964	0.6809
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 12 Robustness Check: Using Seller Incentivized Subsample

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0338***	0.0041***	0.1067***	0.0034***	0.0995***	0.0016***
	(0.0027)	(0.0004)	(0.0031)	(0.0006)	(0.0030)	(0.0005)
Observations	1,613,865	1,613,865	1,613,865	1,613,865	1,613,865	1,613,865
Adjusted R-squared	0.9560	0.3279	0.7973	0.5974	0.8010	0.7194
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 13 Heterogeneous Analyses Over Rating Extremity

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0118***	0.0077***	0.0621***	0.0034***	0.0755***	0.0050***
	(0.0023)	(0.0005)	(0.0045)	(0.0009)	(0.0045)	(0.0008)
Extreme	-0.0199***	-0.0160***	-0.0095***	-0.0025***	-0.0258***	0.0013***
	(0.0008)	(0.0003)	(0.0014)	(0.0003)	(0.0013)	(0.0002)
Extreme×Incentive	0.0023**	0.0045***	0.0339***	0.0045***	0.0318***	0.0018***
	(0.0010)	(0.0004)	(0.0022)	(0.0005)	(0.0021)	(0.0004)
Observations	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808	2,260,808
Adjusted R-squared	0.9572	0.3228	0.7920	0.5957	0.7986	0.7234
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 14 Heterogeneous Analyses Over Search VS. Experience Goods.

	(1)	(2)	(3)	(4)	(5)	(6)
	Syntactic	Semantic	With 10 Aspects		With 20 Aspects	
VARIABLES	Coherence	Coherence	Detailedness I	Detailedness II	Detailedness I	Detailedness II
Incentive	0.0234***	0.0114***	0.0750***	-0.0230***	0.0782***	-0.0093***
	(0.0057)	(0.0019)	(0.0124)	(0.0037)	(0.0114)	(0.0028)
Search	0.0186	-0.0192	0.1791	0.1242***	0.0059	0.0729**
	(0.0236)	(0.0147)	(0.1253)	(0.0338)	(0.1056)	(0.0297)
Search×Incentive	0.0081*	0.0065***	0.0541***	0.0193***	0.0198*	0.0154***
	(0.0047)	(0.0018)	(0.0108)	(0.0035)	(0.0101)	(0.0025)
Observations	213,461	213,461	213,461	213,461	213,461	213,461
Adjusted R-squared	0.9526	0.3344	0.7445	0.2776	0.7733	0.4079
User FE	YES	YES	YES	YES	YES	YES
Product FE	YES	YES	YES	YES	YES	YES
Review Length	YES	YES	YES	YES	YES	YES
Time FE	YES	YES	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES	YES	YES

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 15 Incentive Effect on Review Length Quality

	(1)	(2)	(3)	(4)
VARIABLES	Main Analysis	Analysis With Product/User Experience	Analysis With Verified-Purchase Subsample	Analysis With Seller-Incentivized Subsample
Incentive	-0.0475***	-0.0511***	0.1180***	0.3425***
	(0.0074)	(0.0074)	(0.0093)	(0.0047)
Observations	2,260,808	2,260,808	1,323,335	1,613,865
Adjusted R-squared	0.6556	0.6562	0.6541	0.6981
User FE	YES	YES	YES	YES
Product FE	YES	YES	YES	YES
Review Length	YES	YES	YES	YES
Time FE	YES	YES	YES	YES
Regional Market FE	YES	YES	YES	YES
User Experience		YES		
Product Experience		YES		

Note. The dependent variable across all columns in this table is the log transformation of review length.

Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Table 16 Descriptive Statistics for Continuous Variables

Experiment I

Variable	Obs	Mean	Min	Max	Std. Dev.
Age	1246	35.8138	10	83	12.0504
Syntactic Coherence	1246	-1.29E-09	-4.5602	1.8007	1
Semantic Coherence	1246	0.5425	0.1429	0.9001	0.1164
Detailedness I	1246	2.3938	0	3.5553	0.4846
Detailedness II	1246	1.1004	1.0986	1.6094	0.0261
Review Length	1246	51.1613	9	172	25.0188

Experiment II

Variable	Obs	Mean	Min	Max	Std. Dev.
Age	1080	33.7037	18	73	10.0873
Syntactic Coherence	1080	-3.71E-10	-4.269	1.6405	1
Semantic Coherence	1080	0.5347	0.1742	0.8622	0.1064
Detailedness I	1080	2.4105	0	3.7842	0.5297
Detailedness II	1080	1.1003	1.0986	1.7918	0.0289
Review Length	1080	55.6750	9	187	28.4163

Table 17 Descriptive Statistics for Categorical Variables

Experiment I					
Education		English Proficiency		Gender	
Associate's degree	166	Acceptable	10	Female	773
Bachelor's degree	435	Good	39	Male	462
Graduate degree	172	Native	1,083	Other	11
High school	139	Proficient	114		
Less than high school	8				
Some college	326				
Experiment II					
Education		English		Gender	
Associate's degree	149	Acceptable	14	Female	666
Bachelor's degree	356	Good	43	Male	396
Graduate degree	126	Native	923	Other	18
High school	138	Proficient	100		
Less than high school	6				
Some college	305				

Table 18 Randomization Check

Experiment I			*Experiment II*		
Variables	Chi-Square	P-Value	Variables	Chi-Square	P-value
Age	0.05	0.823	Age	2.0690	0.1503
Gender	2.7003	0.259	Gender	6.9547	0.031
Education	6.3128	0.277	Education	12.5961	0.027
English	2.0501	0.727	English	0.3278	0.988

Table 19 Regression Results Using Experimental Data

	(1)	(2)	(3)	(4)
VARIABLES	Syntactic Coherence	Semantic Coherence	Detailedness I	Detailedness II
		Experiment I		
Incentive	0.0440**	0.0281***	0.0462**	0.0027*
	(0.0174)	(0.0062)	(0.0196)	(0.0015)
Observations	1,246	1,246	1,246	1,246
Adjusted R-squared	0.8030	0.1521	0.5150	0.0106
		Experiment II		
Incentive	0.0317*	0.0362***	0.0473**	0.0032*
	(0.0182)	(0.0062)	(0.0209)	(0.0018)
Observations	1,080	1,080	1,080	1,080
Adjusted R-squared	0.8272	0.1018	0.5823	0.0068
Control for Age	YES	YES	YES	YES
Control for Gender	YES	YES	YES	YES
Control for Education	YES	YES	YES	YES
Control for English proficiency	YES	YES	YES	YES
Control for Review Length	YES	YES	YES	YES

Standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.1

Note. The calculations of Detailedness are based on ten aspects.

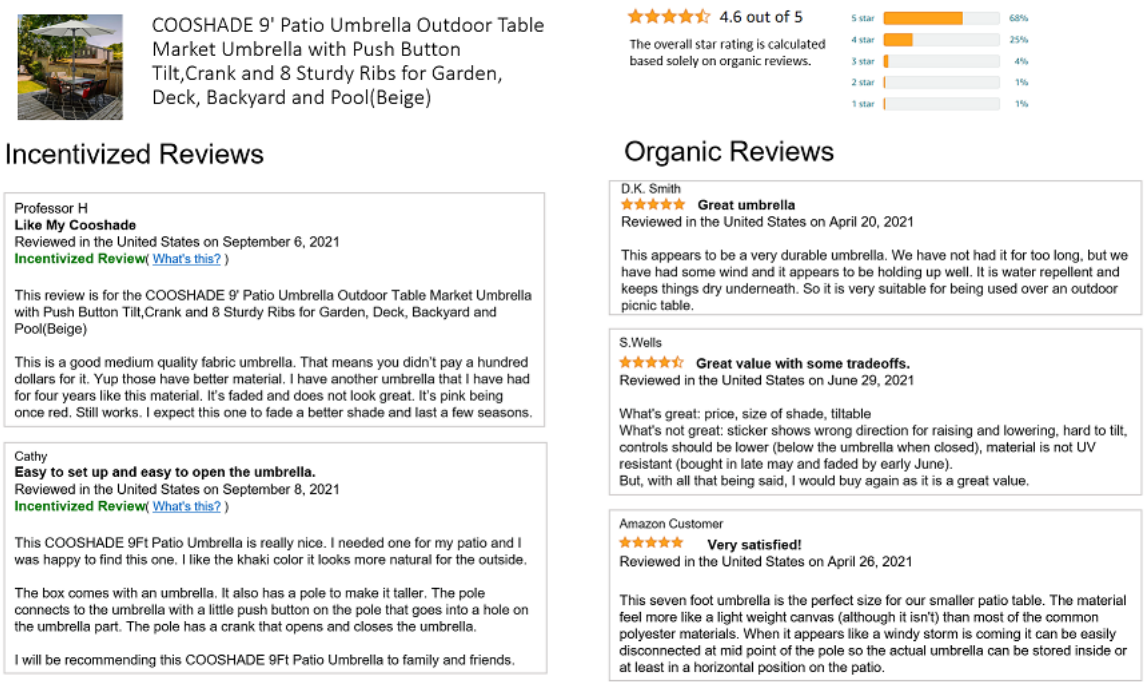
Figure 4 Turk Experimental Flow

Note.

(1) On Page 3, the two questions "Which of the following best describes the Youtuber's attitude?" enquire Youtuber's attitudes towards different attributes of the product.

(2) On Page 4, the descriptions are the messages shown to the treated group. If a participant is allocated to a control group, the corresponding message will be "Although writing review is voluntary, we much appreciate your contribution."

Figure 5 An Illustration of a Future Product Review Interface



References

Angelidis S, Lapata M (2018) Summarizing opinions: Aspect extraction meets sentiment prediction and they are both weakly supervised. *arXiv preprint arXiv:1808.08858* .

Barzilay R, Lapata M (2008) Modeling local coherence: An entity-based approach. *Computational Linguistics* 34(1):1–34.

Burtch G, He Q, Hong Y, Lee D (2019) Peer awards increase user content generation but reduce content novelty. *Available at SSRN 3465879* .

Burtch G, Hong Y, Bapna R, Griskevicius V (2018) Stimulating online reviews by combining financial incentives and social norms. *Management Science* 64(5):2065–2082.

Cabral L, Li L (2015) A dollar for your thoughts: Feedback-conditional rebates on ebay. *Management Science* 61(9):2052–2063.

Caliendo M, Kopeinig S (2008) Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys* 22(1):31–72.

Carenini G, Cheung JCK, Pauls A (2013) Multi-document summarization of evaluative text. *Computational Intelligence* 29(4):545–576.

Chen H, Hu YJ, Huang S (2019) Monetary incentive and stock opinions on social media. *Journal of Management Information Systems* 36(2):391–417.

Clark T, Salaman G (1998) Creating the 'right'impression: towards a dramaturgy of management consultancy. *Service Industries Journal* 18(1):18–38.

Duan Y, Chen C, Huo J (2019) The impact of monetary rewards for online reviews. *Asia Pacific Journal of Marketing and Logistics* 31(5):1486–1515.

Elsner M, Charniak E (2011) Extending the entity grid with entity-specific features. *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 125–129.

Foltz PW, Kintsch W, Landauer TK (1998) The measurement of textual coherence with latent semantic analysis. *Discourse processes* 25(2-3):285–307.

Goffman E, et al. (1978) *The presentation of self in everyday life* (Harmondsworth London).

Grosz BJ, Weinstein S, Joshi AK (1995) Centering: A framework for modeling the local coherence of discourse. *Computational linguistics* 21(2):203–225.

Grove SJ, Fisk RP (1983) The dramaturgy of services exchange: an analytical framework for services marketing. *Emerging perspectives on services marketing* 45–49.

Grove SJ, Fisk RP (1992) The service experience as theater. *ACR North American Advances* .

Hainmueller J (2012) Entropy balancing for causal effects: A multivariate reweighting method to produce balanced samples in observational studies. *Political Analysis* 20(1):25–46.

Halliday M, Hasan R (1976) Cohesion in english. *English Language Series, Longman, London* .

He R, Lee WS, Ng HT, Dahlmeier D (2017) An unsupervised neural attention model for aspect extraction. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 388–397.

Higgins D, Burstein J, Marcu D, Gentile C (2004) Evaluating multiple aspects of coherence in student essays. *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics: HLT-NAACL 2004*, 185–192.

Hsieh G, Kraut RE, Hudson SE (2010) Why pay? exploring how financial incentives are used for question & answer. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 305–314.

Karamanis N, Poesio M, Mellish C, Oberlander J (2004) Evaluating centering-based metrics of coherence for text structuring using a reliably annotated corpus. *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, 391 (Association for Computational Linguistics).

Khern-am nuai W, Kannan K, Ghasemkhani H (2018) Extrinsic versus intrinsic rewards for contributing reviews in an online platform. *Information Systems Research* 29(4):871–892.

Kuang L, Huang N, Hong Y, Yan Z (2019) Spillover effects of financial incentives on non-incentivized user engagement: Evidence from an online knowledge exchange platform. *Journal of Management Information Systems* 36(1):289–320.

Lin Z, Zhang Y, Tan Y (2019) An empirical study of free product sampling and rating bias. *Information Systems Research* 30(1):260–275.

Liu Y, Feng J (2016) Does money talk? the impact of monetary incentives on user-generated content contributions. *working paper* .

McHugh ML (2012) Interrater reliability: the kappa statistic. *Biochemia medica* 22(3):276–282.

Ni J, Li J, McAuley J (2019) Justifying recommendations using distantly-labeled reviews and fine-grained aspects. *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 188–197.

Poesio M, Stevenson R, Eugenio BD, Hitzeman J (2004) Centering: A parametric theory and its instantiations. *Computational linguistics* 30(3):309–363.

Qiao D, Lee SY, Whinston AB, Wei Q (2020) Financial incentives dampen altruism in online prosocial contributions: A study of online reviews. *Information Systems Research* 31(4):1361–1375.

Reiter E, Dale R (2000) Building natural generation systems. *Studies in Natural Language Processing. Cambridge University Press* .

Solomon MR, Surprenant C, Czepiel JA, Gutman EG (1985) A role theory perspective on dyadic interactions: the service encounter. *Journal of marketing* 49(1):99–111.

Stein NL, Glenn CG (1975) An analysis of story comprehension in elementary school children: A test of a schema. .

Stephen A, Bart Y, Du Plessis C, Goncalves D (2012) Does paying for online product reviews pay off? the effects of monetary incentives on content creators and consumers. *NA - Advances in Consumer Research* 228–231.

Sun M, Zhu F (2013) Ad revenue and content commercialization: Evidence from blogs. *Management Science* 59(10):2314–2331.

Sun Y, Dong X, McIntyre S (2017) Motivation of user-generated content: Social connectedness moderates the effects of monetary rewards. *Marketing Science* 36(3):329–337.

Swartz T, Iacobucci D (2000) *Handbook of services marketing and management* (Sage).

Wang J, Ghose A, Ipeirotis P (2012) Bonus, disclosure, and choice: what motivates the creation of high-quality paid reviews? *Proceedings of International Conference on Information Systems* (Association for Information Systems).

Wang RY, Strong DM (1996) Beyond accuracy: What data quality means to data consumers. *Journal of management information systems* 12(4):5–33.

Wang S, A Pavlou P, Gong J (2016) On monetary incentives, online product reviews, and sales. *Proceedings of International Conference on Information Systems* (Association for Information Systems).

Wang X, Sanders GL (2019) For money, and for fun: Exploring the effects of gamification and financial incentives on motivating online review generation. *Proceedings of Americas Conference on Information Systems* (Association for Information Systems).

Wang Y, Sun A, Han J, Liu Y, Zhu X (2018) Sentiment analysis by capsules. *Proceedings of the 2018 world wide web conference*, 1165–1174.

Witte SP, Faigley L (1981) Coherence, cohesion, and writing quality. *College composition and communication* 32(2):189–204.

Woolley K, Sharif MA (2021) Incentives increase relative positivity of review content and enjoyment of review writing. *Journal of Marketing Research* 58(3):539–558.

Yu Y, Khern-am nuai W, Pinsonneault A (Forthcoming) When paying for reviews pays off: The case of performance-contingent monetary rewards. *MIS Quarterly* .