Incentive Provision and Pro-Social Behaviors

Dandan Qiao
Tsinghua University
qiaodd.12@sem.tsinghua.edu.cn

Shun-Yang Lee
University of Texas at Austin
shunyang.lee@utexas.edu

Andrew B. Whinston
University of Texas at Austin
abw@uts.cc.utexas.edu

Qiang Wei
Tsinghua University
weiq@sem.tsinghua.edu.cn

Abstract

Individuals' pro-social behaviors are driven by altruistic and selfish motivations. In this paper we explore how the introduction of external incentives would influence one's pro-social behavior both in the short term and in the long run. Using a large data set on Amazon product reviews, we design a quasi-experimental approach where we combine a propensity score matching (PSM) and a difference-in-differences (DiD) method to empirically study the effect of incentive provision on reviewer's behavior. We apply techniques from linguistics, language processing, and machine learning to propose several novel measures to capture reviews' writing style and quality. We find evidences consistent with crowding-out and overjustification effects. Our study contributes to the understanding of pro-social behavior and sheds light on how incentives would shift individual behavior.

1. Introduction

Human behaviors are fundamentally interesting yet extremely complex. We regularly undergo a complex decision-making process, balance desires and wants, in order to determine our best course of action. The classical Homo Economicus paradigm argues that individuals tend to act selfishly to maximize their own self-interests and private benefits [1, 2]. However, numerous research also contends that people are inherently altruistic and are capable of showing concerns toward others' welfare [3]. Ample evidences of altruistic behaviors have been found in multiple disciplines including neuroscience, biology, psychology, and economics, with examples such as charitable giving and other pro-social behaviors [4, 5, 6, 7].

This mixture of altruism and selfishness can also be observed in the realm of information systems. The sustainability of online content sharing platforms such as YouTube, Flickr, Wiki, Yelp, and even the review community on Amazon relies on users' voluntary contribution. Users on these websites voluntarily devote time and effort to contribute knowledge and information, also known as user-generated content (UGC). While altruism is often credited as the major impetus that drives people's voluntary contribution behavior on these platforms, we also observe more selfish motivations such as the need for self-enhancement, external recognition, reputation gaining, and social connectedness that prompt voluntary contribution online, similar to other pro-social behaviors [8, 9, 10]. These self-gratification and altruistic tendencies together help shape individuals' contribution behaviors [7, 9, 11].

Many have attempted to encourage pro-social behaviors such as blood donation and charitable giving by providing external rewards. Such incentive provision strategies are expected to stimulate individuals' reciprocal emotions [12], and induce them to behave in a way benefiting the incentive provider, such as increasing the contribution level, in return of the favorable treatment. However, abundant studies find that the incentive can actually backfire to crowd out the contributor's original altruistic and intrinsic tendencies for pro-social activities [13, 14, 15,16]. As a consequence, individuals will decrease their contribution or lower their performance in these pro-social activities. Moreover, such negative impacts, as suggested by psychological theory, involves individuals' internal mindset change and hence can persist long enough to generate some overjustification effect [16, 17, 18, 19]. It means that individuals continue to perform in a low level in subsequent activities even when the incentives are no longer present.

The adoption of incentives on those pro-social contribution based information systems are not unusual and begins to attract academic attention in this area. Particularly, Cabral and Li (2015) conducted field experiments on eBay and found that users do show reciprocity by decreasing the likelihood of leaving negative feedback when they receive rebates [20]. Other researchers have empirically studied the temporary crowding-out effect and find that reviewers

URI: http://hdl.handle.net/10125/41838
ISBN: 978-0-9981331-0-2

would decrease the average length of reviews or increase the numerical rating when they receive payments [21, 22, 23, 24]. However, to the best of our knowledge, none of the extant literature studies the incentive's long-term effect on the reviewer's behavior after the incentive has been later removed.

Drawing upon psychological and economic theories, we develop a series of hypotheses on incentives, which may have short-term crowding-out effect as well as long-term overjustification effect, in the context of review contribution. We then empirically test them using a large dataset from Amazon online reviews, the leading e-commerce platform, which allows consumers to voluntarily and pro-socially contribute product reviews. To comprehensively understand the reviewer's behavioral change under incentive provision, we innovatively analyzed the text structure (lexical richness), semantics (topic diversity), and helpfulness (peer evaluation) of their reviews through an integrated linguistics, language processing, and machine learning approach. With these evaluation metrics, we specify a series of fixed-effects models to examine the influence of incentives on reviewers' behaviors both in the short term and the long term. To account for potential endogeneity issues due to reviewers' self-selection, we combine a propensity score matching (PSM) and a difference-in-differences (DiD) approach to conduct our empirical analysis. This enables us to simulate a quasi-experimental environment, which will allow us to estimate the "treatment effect" of incentive provision on the reviewer's behavior.

Our empirical results suggest that the provision of incentives does affect individuals' behaviors, both in the short term and the long term. This external incentive will crowd out the reviewer's initial altruistic and intrinsic motivations. Moreover, the crowding-out effect will carry over to the individual's subsequent reviews, which results in the overjustification effect. Specifically, we observe that the provision of incentives corresponds to an immediate decrease in the helpfulness and the lexical richness of the review, while an increase in the review's topic diversity. Besides the short-term effect, we are also able to analyze reviewers' long-term behavior change induced by incentive provision because our dataset contains a given reviewer's review writing history over time. We find evidence consistent with the long-term overjustification effect where we observe a decline in lexical richness among reviewers not currently receiving incentives but having received incentives in the past. The combination of PSM and DiD allows us to make causal arguments regarding these behavior changes. These results ought to be substantially important for the design of online review platforms and other information communities which depend on pro-social contributions. Managers on these platforms must carefully consider this unintended consequence and decide whether or not the provision of incentives is the best strategy for the long-term development of the platform. In addition, this study advances our understanding of pro-social behavior and the interplay among altruism, selfishness, and incentives. We believe our research should be of interest to academics and practitioners interested in promoting pro-social behaviors, and we hope to raise the awareness of the unintended consequence of the incentive provision on both short- and long- term behavior change.

2. Hypotheses Development

Contribution to online review platforms such as Amazon and Yelp is a pro-social example where users share their opinions and experiences to help others search for relevant information and make purchase decisions. Meanwhile, they might also be intrinsically motivated by the desire for social connectedness, uniqueness, self-enhancement, and reputations [10]. Therefore, users' online review contribution is likely a result of a combination of altruism and selfishness.

However, this mixture behavior resulted from altruism and selfishness is easily influenced by financial incentives [15, 16]. People are inclined to make some attributions about the causes of their own behavior based on the behavior per se and the conditions within which it occurs [17, 25]. When no explicit exogenous factors are available, subjects act as if their behaviors are intrinsically motivated [15, 19, 26, 27]. For example, on the review platform, reviewers are willing to share their opinions on products to help others make purchase decision. They will probably attribute their pro-social contribution to altruistic attitudes, or some other intrinsic motivations such as self-enhancement and enjoyment [7,10]. Either altruistic or these intrinsic motivations are abstract, only facilitating individuals to find some temporary interpretations for their behaviors. Without interventions, individuals may continuously contribute in the pro-social activities under such internal motivations. However, once some exogenous factor which can affect the pro-social contribution is present, the endogenous interpretations will be diluted. This specific external factor will become a salient reason for individuals to explain and support their behaviors [17, 19, 25, 28]. Back to the review platform, if reviewers are provided with some financial incentives (e.g., free products or cash back), which have been validated as a strong impactor on pro-social activities, it will undoubtedly be an attention

focus in this circumstance. The salience of financial incentives will provoke reviewers' introspect on the association of incentives with their behaviors. They hence begin to infer the causes of their behaviors in a new perspective and have an attributional shift in their minds [17, 19]. To this point, the financial incentives seem to give a sufficient and reasonable interpretation for their pro-social contributions compared with the original ambiguous and invisible internal motivations. This attributional shift transfers reviewers from a pro-social context to a monetary framework [28], where they consequently decrease the initial motivations including altruism and some intrinsic interests. Such decrease of altruism and intrinsic motivations induced by financial incentives is psychologically defined as "crowding-out effect" [13, 15, 26]. Our first set of hypotheses concerns the incentive's crowding-out effect on reviewers. Specifically, we suspect that these contributors may be less likely to devote effort in writing, resulting in a decline in the review helpfulness voting and in the quality of the review text itself, as reflected in certain measures of lexical richness, which is an indicator of writing quality. It is also likely that incentivized reviewers might care less about the peer readers and thus pay less attention to the objectivity and accuracy of the review. As a result, the topic diversity covered in the text might become more biased. Therefore, we hypothesize the following:

H1: (Crowding-out Effect) Reviewers who received incentives from the seller would write reviews with biases including: (a) less helpfulness, (b) lower quality as reflected in lexical richness, (c) significantly different topic diversity, compared with reviews given by a reviewer not receiving incentives.

As mentioned above, it's not a simple association of the extrinsic rewards with the pro-social activity but rather the perception of extrinsic incentive salience that would be necessary to produce a crowding-out in initial internal motivations, i.e., altruism and intrinsic satisfactions in pro-social activities. This means that individuals update beliefs about their behavioral attributions when the incentive provision is present. Once incentives are withdrawn, such updated perceptual beliefs will not disappear immediately from individuals who actually have already been overjustified by such incentives [17, 15, 19]. It's hence not surprising that these intangible thoughts derived from previous incentives will continue to influence individuals' subsequent behaviors even when the incentives are no longer present. Therefore, the original motivation to perform the task without incentives can be reduced in the long run, leading to the "overjustification effect" [17, 18, 19]. Consistent evidences of such long-term effects from incentives have been found in different studies. For example, children who received some expected rewards would show less subsequent interest in an attractive drawing activity even when the incentives were removed later [17]. In a day care experiment, some fine policy was enacted to punish parents who were late to pick up children. When the fine rule was abolished, parents still showed up late and the number of late pick-ups even increased [29]. All these examples supported that incentives can produce long-term overjustification effects on individual behaviors even when the incentives are later removed.

In the current context of a review platform, reviewers often post reviews for more than one product. It is possible that certain reviewers would receive some incentives to write a review for a certain product, and at a later point in time write reviews for other products without receiving incentives. According to the overjustification effect, the crowding-out from incentives could still persist and influence the reviewers' behavior in these subsequent reviews, which implies that we might still observe some difference in the topic diversity, and the writing quality of the review text. Therefore, we propose the following hypotheses:

H2: (Overjustification Effect) Reviewers who have received incentives in the past would in later periods write reviews with biases including: (a) less helpfulness, (b) lower quality as reflected in lexical richness, (c) significantly different topic diversity, compared with reviews given by a reviewer not receiving incentives.

3. Research Setting and Measurements

3.1. Amazon Online Review

Amazon, the largest e-commerce platform, allows users to leave product reviews, and the provision of product reviews is usually voluntary: users typically contribute reviews to help other potential buyers without receiving any financial rewards. While review provision can appear to be a purely altruistic behavior, one might be able to establish his/her reputation as a respected reviewer as s/he contributes more helpful reviews on Amazon. Therefore, this review contribution behavior is likely to reflect a combination of altruism as well as one's intrinsic selfish motivations to obtain self-enhancement and respect from others. Interestingly, in 2007 Amazon launched the Amazon Vine Program, through which reviewers are eligible to receive free or discounted sample products in exchange of writing an honest and unbiased product review. The Vine program makes Amazon review an ideal platform for us to study whether or/and how users might change their behavior in response to external financial incentive. Amazon

requires such reviews to explicitly disclose in the review text that a free/discounted product has been received, which allows us to identify reviews that are written when the external financial incentive is present. To distinguish these reviews from others, we label them as "incentivized reviews". In addition, we further label users who have written any incentivized review as "incentivized reviewers". An important fact is that an "incentivized user" can still write "non-incentivized reviews" for other products because the Vine program is conducted on a per product basis.

To explore whether/how the provision of incentive would affect individual behavior, we obtain a rich dataset which contains Amazon's product and review information for selected product categories from 1997 to 2014. [1] Our first task is to identify reviews that were incentivized. Since reviewers have to disclose the receipt of financial incentives in the review text but the actual disclosure texts might vary across reviews. We consider the problem of incentive identification as a text classification task where we classify review texts as either incentivized or non-incentivized, and we use a powerful machine learning method, random forest [30], to identify incentivized reviews: we first recruit human evaluator to annotate a subset of review texts as either incentivized or non-incentivized. We then use the labeled data to grow a random forest with N-gram features as inputs, and this random forest classifier achieves 92.21% accuracy, suggesting that our classifier is satisfactory. This classifier enables us to automatically extract incentivized reviews from our large dataset. Reviewers who post these incentivized reviews are consequently identified as incentivized reviewers. We restrict our attention to incentivized reviewers who have posted more than ten identified incentivized reviews to ensure we are capturing true incentivized reviewers. Following this procedure, we identify a total of 4118 incentivized reviewers, all of whom have written more than 10 incentivized reviews. They collectively contributed a total of 1,165,035 pieces of reviews, of which 191,429 reviews are identified as incentivized. We also identify 1,399,384 reviewers who have never received any incentives but post more than 10 reviews per person. These 4,118 incentivized reviewers and 1,399,384 regular reviewers are the focus of our following empirical analysis.

3.2. Measurements Development

Helpfulness: Amazon implements a voting mechanism for customers to evaluate others' reviews, which allows us to evaluate reviewers' contribution [31]. We collect this information to calculate the helpfulness as ratio of helpful votes among all votes which we use as a measure of review quality. Since reviews with high helpfulness are perceived to be useful in facilitating purchase decisions [32], a high helpful ratio suggests that the review quality is high and that the reviewer has exerted some effort in writing the review.

Lexical Richness: We also analyze the structure of a review text to evaluate the quality of the writing. The linguistic concept, lexical richness, captures the degree to which the writer is using a varied and large vocabulary [33], and it has been shown to be positively related to the quality of written and spoken texts in language studies [34]. Texts that have a high level of lexical richness are more capable of expressing complex ideas, especially when complex ideas are combined or interacting with each other [35]. We therefore include it as a measure of review quality in our analysis. We use a natural language processing technique (POS) to parse the reviews and further adopt an automated lexical analysis tool, Lexical Complexity Analyzer [36] to analyze a review text's lexical richness.

Topical Diversity: Since a comprehensive evaluation of a product is more likely to be helpful, another way to measure review quality is to analyze if a review text discusses a diverse set of topics: reviews that contain a higher number of different topics might be more informative and diagnostic in helping users make their purchase decision. We use a type of topic model, Latent Dirichlet Allocation (LDA) [37], to analyze review texts. This method has been widely used in review processing such as to recommend helpful information and to extract product attributes. Given a specific review, LDA outputs its topic distribution. We then apply the concept of Shannon entropy to calculate the topic diversity as follows: for a given review with a topic distribution of $\mathrm{T} = \{\mathrm{t}_1, \mathrm{t}_2, \ldots, \mathrm{t}_\mathrm{n}\}$, we define topic diversity as

$$TopicDiversity = -\sum_{i=1}^{n} t_i \log(t_i).$$

4. Empirical Analysis

Using the Amazon online review dataset, we construct a series of empirical models to test our hypotheses to study both the short-term and the long-term effects of incentive provision on the reviewer's behavior. Here we design a quasi-experimental environment which combine the propensity score

[1] We thank Dr. Julian McAuley for generously providing this dataset (http://jmcauley.ucsd.edu/data/amazon/)

matching (PSM) technique and the difference-in-differences (DiD) analysis to address the potential self-selection issue.

4.1. Propensity Score Matching

Propensity score matching (PSM) has become a popular *quasi-experimental* method to estimate treatment effects and has been increasingly applied in IS research. Here we apply PSM to construct a "control group" which consists of reviewers who never received any incentives for writing reviews but exhibit very similar review contribution pattern and writing styles compared with the "treatment group"—reviewers who did receive incentives—*before* any incentive was provided. The PSM procedure involves matching a given treatment user with a similar control user based on observable covariates. It is important to note that the *conditional independence assumption* (CIA) has to satisfy for the treatment effect estimate to be valid. This assumption states that the potential outcome of the treatment is independent of the treatment assignment *conditional* on observable covariates. Rosenbaum and Rubin (1983) show that, if CIA holds, then matching based on propensity score—the likelihood of receiving treatment—is sufficient [38]. In our case, we model the propensity of receiving treatment, i.e., incentives, as a function of variables that reflect the reviewer's review writing characteristics *prior* to any incentive provision.

One challenge specific to our research setup is that incentivized-reviewers do not necessarily receive incentive at the same time. In other words, there does not exist an overall "treatment start time" where all treatment users start to receive incentives for review writing. This creates an issue for our analysis because we need to match treated users with control users *prior* to the "treatment start time" for each user, but none of the control users ever received any "treatment" by definition, and hence we do not have a well-defined "treatment start time" for the control users. Our approach to this issue is to conduct a *two-stage* PSM procedure where we first use data from the entire period of time (both prior to treatment and after treatment) to match treatment users with control users who are similar overall. Then, for each control user, we *assign* this control user's "treatment start time" to be the matched treatment user's "treatment start time". Once all control users have been assigned their respective "treatment start time", we are able to conduct our *second-stage* PSM using only data prior to the "treatment start time", and the second-stage PSM gives us the final matches for each treatment user. In the subsequent analysis we also conduct robustness checks to ensure our assignment of "treatment start time" for control users does not qualitatively affect our empirical results. The results of the robustness checks are available upon request.

For both stages of the PSM procedure, we perform a Nearest Neighboring matching (NN matching), which pairs each incentivized reviewer (treatment) with the closest non-incentivized (control) reviewer in terms of their propensity scores. We specify a logistic regression to model each reviewer's probability of being incentivized. Note the variables used in the logistic regression are the review length, numerical star rating, helpful ratio, and number of total votes, all of which are averaged among all reviews written by a given reviewer. We also include measures such as the number of reviews, length of tenure on the platform, and number of product categories this reviewer has reviewed into the regression to reflect additional review characteristics. Using these variables, we employ a stepwise estimation [39] to specify the final propensity score formula. It is important to note that a common support of propensity score is required for PSM to work properly. Therefore, we discard observations which lie outside of the common support region based on the Minima and Maxima comparison suggested by Caliendo and Kopeinig [40]. Finally, we conduct the two-stage PSM procedure using the NN method with replacement.

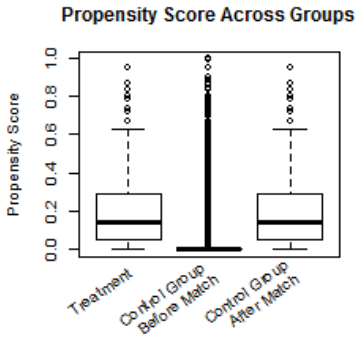

Figure 1. Distribution of propensity score before and after 1st-stage matching

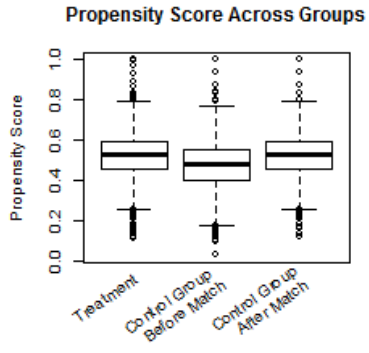

Figure 2. Distribution of propensity score before and after 2nd -stage matching

To ensure that our matching is successful, we plot the distribution of propensity score before and after matching between two groups in Figure 1 and Figure 2. We can see the propensity score distribution of the control group after matching is almost identical to that of the treatment group, which suggests that matching is satisfactory. We further conduct statistical tests and conclude that the distributions of all variables are not significantly different between control and treatment group after each matching. Due to space constraints, statistical test results are available upon request.

4.2. Difference-in-Differences Model

The two-stage propensity score matching procedure allows us to simulate a quasi-experiment by pairing each incentivized reviewer with a similar non-incentivized reviewer based on *observable* covariates. To alleviate the concern regarding selection on *unobservables,* we combine the PSM procedure with a difference-in-differences (DiD) approach, which allows us to also address time-invariant differences in unobservables across the treatment and control group.

Note that we are interested in how the *reviewer's overall behavior* is affected by incentive provision, hence making the results less susceptible to any single extreme review that might bias our effect estimates in unexpected ways. For each treatment reviewer, we first split his/her reviews into those that are written *before* "treatment start time" and those *after* "treatment start time"; for those reviews written *after* the "treatment start time" we further split them into those that have been classified as *incentivized reviews* and those that have been classified as *unincentivized reviews*. We call these 3 splits different *regimes*. Then, for each regime we compute the average for each variable that will appear in the empirical model below, such as *HelpfulRatio, LexicalRichness* and *TopicDiversity*. Therefore, for each treatment user we have 3 regimes: treatment-before-unincentivized, treatment-after-unincentivized, treatment-after-incentivized; similarly, for each control user we will have 2 regimes: control-before-unincentivized and control-after-unincentivized. Finally, we construct a consolidated dataset where each entry is one of the 5 potential regime levels for a given reviewer, with each variable in the entry being the *average* of the variable values across all reviews in this regime.

Our outcomes of interest are the reviewer's behaviors measured by *HelpfulRatio, LexicalRichness*, *TopicDiversity*. Recall that *LexicalRichness and TopicDiversity* are direct measurements of the reviewer's own behavior, while *HelpfulRatio* is the peer evaluation of the review quality. Considering this difference, we therefore specify regime-level models separately for these two types of measurements. For measures that directly reflect the reviewer's behavior, we specify the following regime-level, fixed-effect DiD model:

$$\begin{aligned} DV_{ir} = \alpha_{0i} + \alpha_1 Treat_{ir} + \alpha_2 IncentivizedStatus_{ir} \\ + \alpha_3 Treat_{ir} \\ \cdot IncentivizedStatus_{ir} \\ + \alpha_4 IncentivizedReview_{ir} \\ + \alpha_5 RegimeSpecificVariables_{ir} \\ + \alpha_6 ReviewerCharacteristics_{ir} \\ + \alpha_7 ProductCharacteristics_{ir} \\ + \alpha_8 Time_{ir} + \varepsilon_{ir}, \end{aligned}$$

where i indexes a matched pair of reviewers, and r indexes one of the 5 regime types; $Treat_{ir}$ is the treatment dummy variable which equals 1 when the reviewer is in the 'treatment'-regime, and 0 otherwise; $IncentivizedStatus_{ir}$ is a dummy variable which equals 1 if the observation corresponds to an 'after'-regime, and equals 0 otherwise. In other words, this dummy variable reflects the difference in the before-after dimension in the DiD model. $IncentivizedReview_{ir}$ is a dummy variable which equals 1 if the observation corresponds to an 'incentivized'-regime, and 0 otherwise. $RegimeSpecificVariables_{ir}$ represents a vector of variables which reflect the reviewer's reviewing behavior including *ReviewLength, NumReviews, ActiveMonths* in the current regime. The purpose of including regime-specific variables can be understood as controlling for the reviewer's characteristics in the current regime, so we can account for any potential difference in the reviewer's characteristics across different regimes. In contrast, we also control for $ReviewerCharacteristics_{ir}$, which is a vector reflecting a reviewer's *overall characteristics* as reflected in all regimes, as opposed to his/her characteristics in the current regime. $ProductCharacteristics_{ir}$ denotes the overall characteristics of all products for which the reviewer has evaluated, including *NumCategory, AveragePrice* and *AverageSalesRank.* Finally, $Time_{ir}$ is a year-fixed effect denoting the year when the reviewer first started writing reviews; $\alpha_{0i}'s$ are the pair-specific fixed effects that capture the baseline differences between different matched pairs and enable us to control for unobserved heterogeneity across different pairs of reviewers.

Recall that we define incentivized users as those who have ever written some incentivized reviews, but these incentivized users do not always write incentivized reviews. Therefore, the variable $IncentivizedReview_{ir}$ for a treatment user can be either 1 or 0, depending on whether or not the current regime is incentivized or not. In contrast, a control

user is by definition unincentivized—never written any incentivized reviews—and therefore the variable $IncentivizedReview_{ir}$ for the control user is always 0 by definition. Note that our parameters of interests is the coefficient, α_3, associated with the interaction term, $Treat_{ir} \cdot IncentivizedStatus_{ir}$, which measures the difference in *DV between* the treatment and control group *across* the 'before'- and 'after'-regimes—prior to or after the "treatment start time", when the incentive is *NOT* present. In other words, it can be shown that α_3 reflects the following quantity:

$$\begin{aligned}&\alpha_3\\&= [DV(treatment\_after\_unincentivied)\\&- DV(control\_after\_unincentivized)]\\&- [DV(treatment\_before\_unincentivized)\\&- DV(control\_before\_unincentivized],\end{aligned}$$

Which is precisely the definition of the *overjustification* effect due to incentive provision. Similarly, it can be shown that the crowding-out effect is reflected by the sum of α_3 and α_4, where the sum reflects the differences in *DV* between a 'before-unincentivized' regime and an 'after-incentivized' regime—precisely the definition of the crowding-out effect.

We specify a separate model for *HelpfulRatio* because helpfulness votes are cast by peer users and are an indirect reflection of the reviewer behavior. We include review characteristics variables such as *LexcialRichness, TopicDiversit*, in the model because readers are able to observe these features when reading the review, so their votes might have been affected by these features, either directly or indirectly. Formally, we construct the *HelpfulRatio* model as

$$\begin{aligned}HelpfulRatio_{ir} = {} & \alpha_{0i} + \alpha_1 Treat_{ir}\\&+ \alpha_2 IncentivizedStatus_{ir}\\&+ \alpha_3 Treat_{ir}\\&\cdot IncentivizedStatus_{ir}\\&+ \alpha_4 IncentivizedReview_{ir}\\&+ \alpha_5 RegimeSpecificVariables_{ir}\\&+ \alpha_6 LocalBehaviorVariables_{ir}\\&+ \alpha_7 ReviewerCharacteristics_{ir}\\&+ \alpha_8 GlobalBehaviorVariables_{ir}\\&+ \alpha_9 ProductCharacteristics_{ir}\\&+ \alpha_{10} Time_{ir} + \varepsilon_{ir},\end{aligned}$$

where *HelpfulRatio* is defined as the ratio of helpful votes over total votes; $LocalBehaviorVariables_{ir}$ is a vector containing the average values of the four review characteristics variables in the current regime; $GlobalBehaviorVariables_{ir}$ denotes the average values of the four review characteristics variables' across all regimes for the given user; the rest of the variables are defined similarly as in the former model.

5. Model Estimation and Results

In this section we detail the estimation results of our models specified to study the effect of incentive provision and test our hypotheses on the crowding-out and overjutification effects.

5.1. Helpful Ratio Model

We first examine how the helpfulness level of reviews written by a reviewer is affected by the provision of incentives. The estimation results of our DiD model are reported in column (1) of Table 1. We can see that the coefficient associated with the variable *Treat*IncentivizedStatus* is insignificant, which suggests we do not observe any overjustification effect. A probable reason for the lack of overjustication effect is that helpfulness is an evaluation from peer readers, who may be myopic and have no information on whether a reviewer has ever received financial incentives in the past. Therefore, they evaluate the helpfulness of the reviews only based on the current review text without considering any prior incentives. We are also interested in the sum of the coefficient associated with *IncentivizedReview* and the coefficient associated with the interaction term *Treat*IncentivizedStatus,* which is significantly negative. Recall that the sum of these two coefficients captures the crowding-out effect due to incentive provision. This means that, as the reviewer's altruistic and intrinsic motivations are crowded out by the presence of external incentive, the review helpfulness decreases significantly. This result suggests that the seller's strategy to provide reviewers with incentives can lead to reviewers contributing reviews that are less helpful. Note that the coefficient of *Treat* is insignificant, which indicates there is no significant difference between the treatment and control group prior to incentive provision. This suggests that our PSM procedure successfully matched treat users with comparable control users in terms of review helpfulness before 'treatment start date'. In summary, hypothesis *H1a* is supported while *H2a* is not.

5.2. Lexical Richness Model

Here we consider another quality metric, lexical richness. Recall that a higher level of lexical richness is usually linked with better writing quality since it allows the writer to express more complex ideas. As shown in column (2) of Table 1, the estimation results show that the sum of the coefficient associated with *IncentivizedReview* and the coefficient associated with the interaction term *Treat*IncentivizedStatus* is significantly negative[2], which suggests that there is a significant crowding-out effect due to incentive

[2] That is, 0.0122-0.0138 = -0.016

provision. This leads to a significant decrease in lexical richness, an indicator of review quality. Therefore, hypothesis *H1b* is supported. In addition, the coefficient associated with the interaction term, *Treat*IncentivizedStatus*, is also significantly negative, which suggests there is also a negative overjustification effect, and hypothesis *H2b* is supported. This implies that an earlier incentive would negatively affect the reviewer's subsequent review quality once the incentive has been removed.

5.3. Topic Diversity Model

Next we examine how incentive provision would affect topic diversity, a measure of the review content. As can be seen in column (3) in Table 4, the sum of the coefficients associated with *IncentivizedReview* and *Treat*IncentivizedStatus* turns out to be significantly positive, which implies that the topic diversity is affected by the provision of incentives. In other words, when reviewers are provided with incentives, they would on average discuss more topics in the review texts. Therefore, hypothesis *H1c* is supported. Note that we would not have observed any significant result had the incentive provision had no effect on the review content. On the other hand, the coefficient of the interaction term *Treat*IncentivizedStatus* is insignificant, and hence hypothesis *H2c* is not supported. As described above, topic diversity is measured through LDA, which summarizes the topics based on a corpus, i.e., the entire set of review texts in our situation. The aggregation on the huge dataset may lead to information loss of the fine-grained topics specific to each review text. We plan to explore different metrics in order to improve the evaluation of information diversity that each reviewer contributes in his/her review texts.

Table 1. Estimation results of DiD model

Variables	**(1) Helpful Ratio**	**(2) Lexical Richness**	**(3) Topic Diversity**
Treat	-0.0579 (0.0444)	-0.0085 (0.0068)	0.0070 (0.0054)
IncentivizedStatus	-0.3370*** (0.0459)	0.0062 (0.0042)	-0.003 (0.0036)
IncentivizedReview	-0.0812*** (0.0260)	0.0122*** (0.0028)	0.0048** (0.0021)
Treat*IncentivizedStatus	0.0487 (0.0508)	-0.0138** (0.0057)	-0.0046 (0.0049)
ReviewLength	0.2390*** (0.0494)	-0.0529*** (0.0094)	0.12*** (0.0062)
NumReviews	-0.0641*** (0.0150)	0.0079*** (0.0017)	0.0061*** (0.0016)
ActiveMonths	8.88E-4* (5.30E-4)	-2.79E-5 (6.63E-5)	-1.3E-4*** (4.97E-5)
TotalVote	0.0146*** (4.82E-3)		
Entropy	0.0582 (0.0777)		
LexicalRichness	0.6760 (0.6330)		
Constant	-3.7970*** (0.5890)	0.4300*** (0.1310)	0.9420*** (0.0772)
ReviewerCharacteristics	YES	YES	YES
ProductCharacteristics	YES	YES	YES
Time Effects	YES	YES	YES
Fixed Effects	YES	YES	YES
Adjusted R-squared	0.459	0.426	0.496

6. Discussion and Conclusion

Individuals' pro-social behaviors are a result of altruistic and selfish motivations. Researchers from various disciplines have started to examine how these motivations interact with each other. A particularly interesting question is how external incentives would influence individuals' pro-social behavior. A better understanding of this issue will be beneficial for academics and practitioners hoping to promote pro-social behaviors and improve social welfare. Pro-social behaviors are also an integral part of many popular information systems, such as YouTube, Wikipedia, Yelp, and even the Amazon review platform. These platforms rely heavily on user's voluntary contribution in the form of user-generated content to sustain their growth. Some platforms have tried to provide external incentives in the hope that users will increase their contribution level in response to external incentives. Therefore, it is important that we understand the effect and implication of such incentive provision on users' contribution behavior.

Using a large data set of Amazon product reviews, we designed a quasi-experimental setup where we combined the propensity score matching (PSM) method with a difference-in-differences (DiD) approach. To capture the review's writing style and quality, we applied techniques from linguistics, language processing, and machine learning to innovate novel measures which reflect the structure and semantics of review texts. We then proposed a comprehensive empirical framework to analyze reviewers' behavior changes using these innovative measures. With these measures, we proposed and estimated a series of fixed-effect difference-in-differences models to examine how incentive provision influences reviewer's contribution behavior.

Our analysis uncovered the short-term and the long-term effects, i.e. crowding-out effect and overjustification effect, respectively, of incentive provision on the reviewer's contribution behavior. We argued that these changes can be explained by the fact that the external incentives might have implicitly shifted an individual's decision-making context from a pro-social environment to an incentive-based environment. The salience of incentives then greatly reduced the person's original altruistic and intrinsic motivations. Our results also suggested that we can observe a long-term change in behavior: this is because once a reviewer has been given some incentives, his/her mindset change can persist over time to also affect his/her future behaviors, even if the incentive has been removed. In summary, these results imply that the platform and the sellers should carefully evaluate whether or not an introduction of incentives will be beneficial for the long-term development of the platform, since it is likely to crowd out users' initial altruistic and intrinsic motivations and also lead to a long-term overjustification effect. A careless provision of incentives might turn out to be harmful for the platform in the long run and might discourage users' voluntary pro-social behaviors.

Our research extends the review literature by exploring the short-term and long-term effects of incentives on altruistically and intrinsically motivated behaviors, and we believe our research will be of interest to academics and practitioners hoping to promote pro-social behaviors on information systems platforms and beyond. Although our research is not without limitations, we hope this paper can serve as a starting point and encourage more researchers to study the issue of voluntary and pro-social behaviors, as well as how external incentives would affect the short-term and long-term user behavior, perhaps through an experimental setting so as to further strengthen the causal arguments presented in this paper.

7. References

[1] Margolis, H. 1984. Selfishness, Altruism, and Rationality, University of Chicago Press.

[2] Miller, D. T. 1999. "The Norm of Self-Interest.," American Psychologist (54:12), American Psychological Association, p. 1053-1060.

[3] Hoffman, M. L. 1981. "Is Altruism Part of Human Nature?" Journal of Personality and Social Psychology (40:1), pp. 121–137 (doi: 10.1037/0022-3514.40.1.121).

[4] Becker, G. S. 1976. "Altruism, Egoism, and Genetic Fitness: Economics and Sociobiology," Journal of economic Literature (14:3), pp. 817–826.

[5] Harbaugh, W. T., Mayr, U., and Burghart, D. R. 2007. "Neural Responses to Taxation and Voluntary Giving Reveal Motives for Charitable Donations.," Science (316:5831), pp. 1622–1625 (doi: 10.1126/science.1140738).

[6] Krebs, D. L. 1970. "Altruism: An Examination of the Concept and a Review of the Literature." Psychological Bulletin (73:4), pp. 258–302 (doi: 10.1037/h0028987).

[7] Sundaram, D. S., and Hills, B. 1998. "Word-of-Mouth Communications : A Motivational Analysis," Advances in Consumer Research (25), pp. 527–531.

[8] Bierhoff, H. W. 2002. Pro-Social Behaviour, New York: Psychology Press.

[9] Daugherty, T. 2008. "Exploring Consumer Motivations for Creating User-Generated Content." Journal of Interactive Advertising. (8:2), pp. 1–24.

[10] Hennig-Thurau, T., Gwinner, K. P., Walsh, G., and Gremler, D. D. 2004. "Electronic Word-of-Mouth Via Consumer-Opinion Platforms: What Motivates Consumers

to Articulate Themselves on the Internet?" Journal of Interactive Marketing (18:1), Elsevier, pp. 38–52.

[11] Kangas, O. E. 1997. "Self-Interest and the Common Good: the Impact of Norms, Selfishness and Context in Social Policy Opinions," The Journal of Socio-Economics (26:5), Elsevier, pp. 475–494.

[12] Falk, A., and Fischbacher, U. 2006. "A Theory of Reciprocity," Games and Economic Behavior (54:2), pp. 293–315 (doi: 10.1016/j.geb.2005.03.001).

[13] Deci, E. L. 1971. "Effects of Externally Mediated Rewards on Intrinsic Motivation," Journal of Personality and Social Psychology (18:1), American Psychological Association, p. 105.

[14] Bénabou, R., and Tirole, J. 2006. "Incentives and Pro-Social Behavior," The American Economic Review (96:5), American Economic Association, pp. 1652–1678.

[15] Gneezy, U., Meier, S., and Rey-Biel, P. 2011. "When and Why Incentives (Don't) Work to Modify Behavior," The Journal of Economic Perspectives, (25:4), pp. 191-209.

[16] Meier, S., 2006. "A Survey of Economic Theories and Field Evidence on Pro-Social Behavior".

[17] Lepper, M. R., Greene, D., and Nisbett, R. E. 1973. "Undermining Children's Intrinsic Interest with Extrinsic Reward: A Test of the 'Overjustification' Hypothesis," Journal of Personality and Social Psychology (28:1), pp. 129–137 (doi: 10.1037/h0035519).

[18] Tang, S.-H., and Hall, V. C. 1995. "The Overjustification Effect: A Meta-Analysis," Applied Cognitive Psychology, pp. 365–404.

[19] Lepper, M.R. and Greene, D. eds., 2015. "The Hidden Costs of Reward: New Perspectives on the Psychology of Human Motivation". Psychology Press.

[20] Cabral, L., and Li, L. (Ivy). 2015. "A Dollar for Your Thoughts: Feedback-Conditional Rebates on eBay," Management Science (61:9), pp. 2052–2063.

[21] Khern-am-nuai, W., and Karthik, K. 2014. "Extrinsic versus Intrinsic Rewards to Participate in a Crowd Context: An Analysis of a Review Platform," Working paper.

[22] Pavlou, P., and Wang, S. 2015. "How do Monetary Incentives Affect Online Product Reviews and Sales?" in Proceedings of the 21th Americas' Conference on Information Systems, Puerto Rico.

[23] Rabin, M. 1993. "Incorporating Fairness into Game Theory and Economics," The American Economic Review, pp. 1281–1302.

[24] Wang, J., Ghose, A., and Ipeirotis, P. "Bonus, Disclosure, and Choice: What Motivates the Creation of High-Quality Paid Reviews?" in Porceedings of 33rd International Conference on Information Systems, Orlando, FL.

[25] Bem, D. J. 1972. "Self-Perception Theory". In Advances in Experimental Social Psychology, L. Berkowitz (eds.), New York: Academic Press, pp.1-62.

[26] Gagné, M., and Deci, E. L. 2005. "Self‐Determination Theory and Work Motivation," Journal of Organizational Behavior (26:4), Wiley Online Library, pp. 331‐362.

[27] Fehr, E., and Falk, A. 2002. "Psychological Foundations of Incentives," European Economic Review (46:4), pp. 687–724.

[28] Heyman, J., & Ariely, D. 2004. Effort for Payment a Tale of Two Markets. Psychological Science, (15:11), pp. 787-793.

[29] Gneezy, U. and Rustichini, A., 2000. "Fine Is a Price", a. J. Legal Stud., 29, p.1.

[30] Breiman, L. 2001. "Random Forests," Machine Learning (45:1), pp. 5–32 (doi: 10.1023/A:1010933404324).

[31] Mudambi, S. M., and Schuff, D. 2010. "What Makes A Helpful Review? A Study of Customer Reviews on Amazon. Com," MIS Quarterly (34:1), pp. 185–200.

[32] Chevalier, J. a, and Mayzlin, D. 2006. "The Effect of Word of Mouth on Sales: Online Book Reviews," Journal of Marketing Research (43:3), pp. 345–354.

[33] Laufer, B., and Nation, P. 1995. "Vocabulary Size and Use: Lexical Richness in L2 Written Production," Applied Linguistics (16:3), Am Assoc Appl Ling, pp. 307–322.

[34] Crowhurst, M. 1983. "Syntactic Complexity and Writing Quality: A Review," Canadian Journal of Education Canadienne De L'education (8:1), pp.1–16.

[35] Beers, S. F., and Nagy, W. E. 2009. "Syntactic Complexity as a Predictor of Adolescent Writing Quality: Which Measures? Which Genre?" Reading and Writing (22:2), pp. 185–200 (doi: 10.1007/s11145-007-9107-5).

[36] Lu, X. 2012. "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives," Modern Language Journal (96:2), pp. 190–208.

[37] Blei, D. M., Ng, A. Y., and Jordan, M. I. 2003. "Latent Dirichlet Allocation," Journal of Machine Learning Research (3:4-5), pp. 993–1022.

[38] Rosenbaum, P. R., and Rubin, D. B. 1983. "The Central Role of the Propensity Score in Observational Studies for Causal Effects," Biometrika (70:1), Biometrika Trust, pp. 41–55.

[39] Rosenbaum, P. R., and Rubin, D. B. 1984. "Comment: Estimating the Effects Caused by Treatments," Journal of the American Statistical Association (79:385), Taylor & Francis Group, pp. 26–28.

[40] Caliendo, M., and Kopeinig, S. 2008. "Some Practical Guidance for the Implementation of Propensity Score Matching," Journal of Economic Surveys (22:1), Wiley Online Library, pp. 3-72.