Stimulating Online Reviews by Combining Financial Incentives and Social Norms

Gordon Burtch[*], Yili Hong[**], Ravi Bapna[*] & Vladas Griskevicius[+]

Abstract

In hopes of motivating consumers to provide larger volumes of useful reviews, many retailers offer financial incentives. Here, we explore an alternative approach, social norms, wherein we inform people about the volume of reviews authored by peers. We test the effectiveness of using financial incentives, social norms, and a combination of both strategies. In two randomized experiments, one in the field conducted in partnership with a large online clothing retailer based in China, and a second on Amazon Mechanical Turk, we compare the effectiveness of each strategy at stimulating online reviews in larger numbers and of greater length. We find that financial incentives are more effective at inducing larger volumes of reviews, but the reviews that result are not particularly lengthy, whereas social norms have a greater effect on the length of reviews. Importantly, we show that the combination of financial incentives and social norms yields the greatest overall benefit, motivating reviews in greater numbers and of greater length. We further assess the treatment induced self-selection and sentiment bias by triangulating the experimental results with findings from an observational study.

Keywords: randomized experiment, social norms, financial incentives, online reviews, intrinsic motivation

[*] Information & Decision Sciences Department, Carlson School of Management, University of Minnesota

[**] IS Department, WP Carey School of Business, Arizona State University

[+] Marketing Department, Carlson School of Management, University of Minnesota

Stimulating Online Reviews by Combining Financial Incentives and Social Norms

1. Introduction

Online reviews can serve as an excellent source of information for consumers to learn about peer opinions regarding various products and services (Dellarocas 2003). Although subject to certain biases (Luca 2016), online reviews are particularly important in online markets, which are characterized by a great deal of information asymmetry (Dimoka et al. 2012). Unfortunately, like many voluntarily provided public goods (Gallus 2015), online reviews may be acutely under-provisioned (Avery et al. 1999; Anderson 1998; Levi et al. 2012). And, when consumers do provide reviews, they are often brief, limiting their helpfulness to other consumers (Cao et al. 2011; Liu et al. 2007; Mudambi and Schuff 2010). To address this problem, many retailers employ strategies intended to boost the volume and length of reviews, most commonly by offering consumers a small financial incentive in exchange for a review.

However, using financial incentives to solicit online reviews has some drawbacks. For example, a great deal of research has highlighted that offering individuals payments can undermine their intrinsic motivation to perform a task (Deci et al. 1999). This suggests that paying consumers for feedback may lead to a reduction in the effort they exert, resulting in short, uninformative reviews. Paying for feedback may also unduly bias opinion, leading consumers to write more positive reviews (Khern-am-nuai and Kannan 2014). Further, consumers tend to react negatively if they learn that a review author was paid for his or her opinion, discounting the review and inferring negative product quality (Avery et al. 1999; Stephen et al. 2012).

Given the multiple downsides of offering payment for writing reviews, here we consider the following questions: what are alternative ways to stimulate consumers to write more online reviews, and what are alternative ways to stimulate consumers to write lengthier online reviews? First, we consider the use of financial incentives to generate online reviews, a relatively common strategy amongst industry practitioners (Cabral and Li 2015; Fradkin et al. 2016; Stephen et al. 2012; Wang et al. 2012), wherein the retailer offers consumers a small payment in exchange for providing a review[1]. Second, we consider an

[1] When we refer to the use of financial incentives in this paper, we mean the solicitation of truthful consumer evaluations in exchange for payment, regardless of consumer sentiment. That is, we are not referring to payment in exchange for strictly positive reviews or fake reviews.

alternative approach, using social norms to stimulate reviews (Chen et al. 2010), wherein we provide a consumer with information about the volume of reviews recently authored by his or her peers. For example, a social norms approach might inform consumers that in the last month, 2,345 shoppers provided an online review for a given retailer. Providing normative information is essentially costless, and a large body of literature in psychology suggests that social norms are an effective way to encourage desirable behavior (Gerber and Rogers 2009; Ferraro and Price 2013; Allcott 2011; Goldstein et al. 2008). Finally, we test the effectiveness of using a combination of social norms and financial incentives. This influence strategy combines a small monetary payment with information about the behavior of peers into a single, joint-intervention.

Our primary goal in this work is to go beyond pure financial incentives by considering alternative approaches that businesses and review aggregation platforms might employ to stimulate online reviews. By going beyond simply designing a better payment scheme, we seek to identify effective interventions that can overcome some or all of the limitations of financial incentives, which may substitute or complement existing approaches that retailers make use of today.

We employ a multi-methodological research design, testing our questions via two randomized experiments and an econometric analysis of online reviews from Amazon.com, which together enable the estimation of causal effects (Aral and Walker 2011) and help to triangulate a number of interesting findings. We conduct one experiment in the field, in partnership with an online clothing retailer based in China, and a second on Amazon Mechanical Turk (AMT). In each experiment, we solicited consumers to provide online reviews by offering either a financial incentive, supplying them with normative information regarding the number of others who had written reviews, or using a combination of the two. We then assessed the effectiveness of these approaches in terms of two different outcomes: (1) *review volume*: how effective were they in motivating minimal effort by increasing the proportion of consumers who authored a review, and (2) *review length*: how effective were they in motivating more intense effort by writing reviews of greater length.

Our results suggest that financial incentives and social norms are differentially effective in motivating review volumes versus length. In both experiments, we found that financial incentives were more effective at motivating larger volumes of online reviews; people were more willing to exert at least minimal effort to write a review when they were promised a small payment. However, this pattern was reversed when it came to the length of reviews, where social norms were more effective. People wrote longer reviews when they were informed that many other people had written reviews.

Perhaps most interestingly, we find in both experiments that the combined application of financial incentives and social norms delivers the greatest overall benefit, stimulating reviews in both greater volume and of greater length. The effects we observe are economically significant; the joint application of payment and a social norm in our field experiment nearly tripled the volume of consumers who authored a review[2] *and* the reviews that resulted were approximately 50% longer[3]. Our findings therefore suggest that relying on financial incentives or on social norms alone to stimulate reviews is sub-optimal. To our knowledge, our research is the first to consider the effects of using both social norms and financial incentives together to encourage desirable behaviors.

We also assess i) treatment-induced sentiment bias in the reviews consumers author, as well as ii) the relative roles of self-selection and changes in intrinsic motivation in response to our treatments to induce reviews. Regarding the former, although we observe no treatment-specific differences in consumer sentiment in our experiments, our large-scale analysis of Amazon reviews does indicate that financial incentives lead to more positive reviews in that setting. Regarding the latter, although our analyses do not enable us to draw strong conclusions, we do observe some evidence which suggests that our results are attributable to some combination of self-selection *and* behavioral change on the part of subjects.

This research contributes to the literature on online reviews by comparing the effectiveness of using the common approach of financial incentives to stimulate online reviews with the less common alternative of social norms. We also contribute to the broader literature on social norms by considering the distinction between the propensity of individuals to engage in a behavior and the intensity of engagement by a given individual. Whereas the vast majority of past work on social norms has considered behaviors that are singular in nature – that is, where the measured outcomes jointly reflect some combination of volume of participants and intensity of participation – we consider a behavior that includes two sequential decisions: agreeing to participate in the task followed by performing the task. This allows us to theorize and disentangle the effects of financial incentives and social norms on each outcome. This is important, because our findings suggest that the effects of financial incentives and social norms on each outcome are asymmetrical.

In the next section, we review prior work on the use of financial incentives and social norms to motivate behavior. We then sequentially present our two experiments, detailing the research contexts, experimental

---

[2] Compared to a baseline proportion of 4% in the control condition in which no intervention was used.

[3] Compared to a baseline average length of ~52 English characters, or 10 words, in the control condition.

designs, and empirical results. Finally, we offer an interpretation and discussion of our findings, discuss the limitations of our work and suggest a number of avenues for future work.

2. Hypothesis Development

2.1. Financial Incentives

One way to motivate behavior relies on offering financial incentives for desired actions. Research on this approach dates back more than 60 years (e.g., Barnes 1949; Jacques et al. 1951). Economic theory holds that rational individuals are utility-driven, meaning that financial incentives should matter for individuals' behavior. A number of studies have empirically verified this prediction. For example, Volpp (2009) showed that paying individuals to quit smoking increased their likelihood of doing so, and Fryer (2010) found that students could be induced to attend school more regularly with the promise of financial compensation.

Consistent with the expected impact of financial incentives, multiple studies have shown that financial incentives are effective at stimulating behavior online. For example, experimental work has found that financial incentives are effective in motivating people to write reviews on AirBNB.com (Fradkin et al. 2016), provide feedback on eBay (Cabral and Li 2015), and provide reviews for BestBuy (Khern-am-nuai and Kannan 2014). It is important to note that the financial incentives offered in most such studies have generally been quite small (with the exception of Fradkin et al. (2016), who offered $25 in AirBNB credit). Cabral and Li's (2015) paid rebates to subjects of just $1 or $2. Similarly, Khern-am-nuai and Kannan (2014) considered BestBuy's offer of 25 BestBuy reward points in exchange for each review, with a monetary value of $0.50.

Bearing in mind that one of our experiments is situated in the context of AMT, it is useful to consider those studies that have specifically examined how financial incentives influence the supply of labor in that context. For example, Horton and Chilton (2010) asked Turkers to click on a pair of rectangles in a specific order, offering them payment for each completed series of clicks. The authors observed that a larger volume of tasks were completed when greater payment was offered. Mason and Watts (2009) conducted a similar experiment, asking Turkers to complete "ordering" tasks, in which they were required to arrange images into a particular sequence. The authors observed that more tasks tended to be completed when pay was higher. Studies in this space have also considered relatively small financial incentives. Mason and Watts (2009) paid less than $0.10 per task, while Horton and Chilton (2010) paid participants according to a concave function of the number of tasks completed – completing 5 tasks

earned a Turker $0.29 cents, and completing 25 tasks earned a Turker $0.82. Taken together, studies in both AMT and other types of settings have shown that offering small financial incentives can motivate people to engage in a desired behavior. Based on this past work, we hypothesize the following:

H1: Offering financial incentives will lead to an increase in the volume of reviews that are provided compared to simply asking.

2.2. Social Norms

A different way to motivate behavior relies on providing social norms. Social norms refer to the prevalence of a behavior in a relevant population, such as the number of people who have already written reviews. This type of social norm is known as a *descriptive social norm* (Cialdini et al. 1991). Social norms have been shown to be effective in a wide range of contexts, from motivating voter turnout (Gerber and Rogers 2009), to encouraging the reuse of hotel towels (Goldstein et al. 2008) to reducing energy consumption (Allcott 2011; Nolan et al. 2008; Schultz et al. 2007), reducing water use (Ferraro and Price 2013), and increasing consumption of healthy foods (Robinson et al. 2014). For example, Robinson, Fleming, and Higgs (2014) tested how social norms influenced the consumption of fruit and vegetables. They exposed people to social norm-based messages indicating the eating behavior of others, and found that those people ate more fruit and vegetables when they were led to believe that their peers had eaten a large amount of fruit and vegetables.

Social norms influence behavior because seeing what others have done provides information about what is socially "normal" in a given context. The greater the number of people who respond to the same situation in the same way, the more people will perceive the behavior to be correct (Thibaut and Kelley 1959). People therefore use social normative information to determine the most appropriate course of action in a given situation (Cialdini and Trost 1998; Cialdini and Goldstein 2004).

To our knowledge, only one study has explored the use of social norms to stimulate the production of online reviews. Chen et al. (2010) conducted an experiment on a movie reviewing website, *MovieLens*, wherein they inform a random set of subjects via email about the median number of reviews recently authored by their peers. They find evidence that this approach increased rates of reviewing amongst treated subjects, on average. Taken together, past findings give us reason to believe that social norms can have a positive influence on the production of online reviews. Formally:

H2: Providing social norms will lead to an increase in the volume of online reviews that are provided compared to simply asking.

2.3. Stimulating Volume vs. Length of Reviews

The central aim of the current research is to test the effectiveness of providing financial incentives, social norms, and the combination of the two in order to stimulate a greater number of longer reviews. Note that each influence strategy could operate on two different aspects of reviewing behavior. On the one hand, we might simply seek to persuade more consumers to submit a review, increasing review volumes. However, persuading a person to write a review does not imply that he or she will necessarily invest the effort to write a lengthy, informative review. Indeed, the vast majority of online reviews are brief and lack useful information (Cao et al. 2011; Liu et al. 2007; Mudambi and Schuff 2010). Thus, the generation of *lengthy* reviews involves motivating a second and critical aspect of behavior: after a person has decided to write a review, he or she must also be persuaded to expend additional effort to write a longer review. An effective strategy for stimulating lengthy online reviews, which are expected to be more helpful to other consumers, must motivate people to both choose to write a review and choose to invest effort in doing so.

The distinction between review volume, which depends simply on choosing to participate (minimal effort), and review length, which depends on intensity of effort, is important because there is reason to believe that financial incentives and social norms might operate differentially on each outcome. When people are provided with a financial incentive to perform a behavior, they are likely to perform that behavior for extrinsic reasons rather than intrinsic reasons (Heyman and Ariely 2004). Offering a financial incentive to write a review is likely to lead people to write the review because they seek to receive the financial reward rather than because of some intrinsic desire to be helpful. The presence of financial incentives therefore shifts people to an effort-for-payment mindset, increasing the probability that they will provide the minimal effort that is warranted, given the level of payment (Heyman and Ariely 2004).

A large body of research indicates that offering financial incentives can change the nature of an individual's task performance by undermining their intrinsic motivation (e.g., Frey 1994; Deci et al. 1999; Jenkins et al. 1998). This suggests that paying people to write a review is likely to undermine people's intrinsic motivation to expend much effort on writing reviews. Moreover, offering financial incentives might elicit reviews specifically from the type of people who lack a pre-existing motivation to write reviews. Sensitivity to financial rewards is a characteristic of individuals that remains relatively stable over time (Rick et al. 2008) and is predictive of selfish behavior in social dilemmas (Seuntjens et al. 2015). Accordingly, individuals who are particularly attracted by the presence of the financial incentive may be predisposed to exert little effort in performing the task, producing short reviews. Taken together,

this suggests that although the common approach of offering financial incentives might be effective at motivating many individuals to write a review, those reviews are likely to be short and relatively uninformative because the people writing them will expend only the minimal effort needed to obtain the reward.

Indeed, research dealing with payment for online task performance shows this exact pattern. Although multiple studies have shown that offering small financial incentives can motivate individuals to complete higher volumes of tasks (Mason and Watts 2009), offering payments does not increase the intensity of effort that individuals dedicate in any given task. For example, Wang et al. (2012) found that offering a financial incentive of the sort we consider here could induce more Turkers to write reviews, but that it had no impact on the quality of those reviews, suggesting no differences in effort intensity under payment. Stephen et al. (2012) likewise found that payment had no effect on the intensity of effort that subjects put into writing evaluations. Finally, Khern-am-nuai and Kannan (2014) observed that, even though consumers began to author larger volumes of reviews following BestBuy's introduction of redeemable reward points, the average length of reviews also declined.

Because online reviews must be sufficiently lengthy to convey meaningful information, and because financial incentives appear to undermine individuals' intrinsic motivation to write longer reviews, we consider whether providing social norms might help fix this problem. Whereas the presence of a financial incentive provides people with an explicit extrinsic reason for why they engaged in a particular behavior ("I wrote the review to receive money"), social norms are more closely linked to intrinsic than extrinsic motivation (Henrich et al. 2006). In fact, people specifically do not view social norms as an extrinsic driver of their behavior (Nolan et al. 2008). Taken together, when people are motivated to write a review as a result of receiving normative information, they likely experience intrinsic motivation in doing so.

The provision of normative information, as with financial incentives, has the potential to induce selection effects, attracting individuals who are predisposed to exert greater levels of effort in the task. As with greed and sensitivity to money, past research has found that a tendency toward altruism and pro-social behavior is a stable trait of the individual (Brief and Motowidlo 1986). This line of reasoning indicates that those individuals who are most likely to be motivated by normative information are those individuals who might also be predisposed toward contributing to the public good. If so, such individuals might also write lengthier and more useful online reviews for the benefit of other consumers. Taken together, the provision of a social norm might result in lengthier online reviews because it may stimulate subjects' intrinsic motivation and it may induce participation by individuals who are predisposed to help others.

This suggests that social norms are likely to be most effective at stimulating lengthy online reviews. Formally:

H3: Providing social norms will lead to an increase in the length of online reviews that are provided compared to providing financial incentives or simply asking.

2.4. Simultaneously Stimulating Volume and Length of Reviews

The ultimate goal of the current research is to identify an influence strategy that motivates both larger volumes of reviews and lengthier reviews. Considering the discussion thus far, financial incentives or social norms alone may be sub-optimal at achieving this goal, especially because financial incentives might undermine the motivation to exert more than minimal effort. We, therefore, consider a third approach: the combined application of social norms and financial incentives. Although no prior work has, to our knowledge, sought to combine social norms and financial incentives to motivate behavior, there is reason to believe that the combined approach might be superior to either approach alone.

We believe that the key to the effectiveness of a combined approach lies in using financial incentives to increase an individual consumer's likelihood of writing a review *without* undermining the intensity of effort that each consumer exerts when writing a review. Evidence for this possibility comes from research in child psychology, which shows that it is possible to circumvent the undermining effects of extrinsic rewards. Cialdini et al. (1998) tested how promising children an extrinsic reward influenced their desire to practice writing skills. Consistent with the classic undermining effect, they found that although promising a reward motivated people to practice writing, the reward undermined children's intrinsic motivation to practice writing when the children were no longer being rewarded for the behavior. However, the research found that, despite the promise of an extrinsic reward for participating, children's intrinsic motivation to practice writing remained high if the kids were subsequently led to believe that they were the sort of children who would want to write well. When the children could attribute their behavior to an internal reason, rather than to an extrinsic reward, they continued to be intrinsically motivated to expend effort on the task. In the same vein, Hennessey and Zbikowski (1993) reported that if children were taught to focus on their own interests as their primary reason for learning, and to treat external incentives as secondary, they were more likely to maintain intrinsic motivation.

We consider the possibility that combining financial incentives with social norms can serve to undermine the undermining effect of external rewards (Cialdini et al. 1998). We hypothesize that the presence of information about the social norm may enable people to rationalize their decision to write a review as one

of goodwill or a personal desire to do what is appropriate, rather than as one of effort for payment. This means that the presence of a financial incentive would serve to motivate people to write the review, but the presence of the social norm would provide people a reason to believe that they chose to write a review for some intrinsic reason rather than solely for financial gain. This leads to our final hypothesis:

> ***H4: Providing social norms and financial incentives, together, will lead to the greatest volume and length of reviews, in tandem.***

3. Empirical Approach

3.1. Research Design

To test our four hypotheses, we first conduct two randomized experiments, one in the field and one on AMT. In each experiment, we compare the effectiveness of providing consumers with (1) a financial incentive, (2) a social norm, (3) a combination of a financial incentive and a social norm, and (4) simply asking them to write a review (our control). To assess effectiveness, we measured how these approaches influence our three outcomes of interest: volume of reviews, length of reviews, and a combination of volume and length.

We evaluate H1 (financial incentives will lead to an increase in the volume of reviews) by comparing the volume of reviews produced in our financial incentive condition with that in the control condition, and we evaluate H2 (social norms will lead to an increase in the volume of reviews) in a similar fashion, comparing the social norm condition with control. To evaluate our third hypothesis, H3 (social norms will lead to an increase in the length of reviews compared to financial incentives or simply asking), we compare the length of reviews authored in our social norm condition with those authored in our control and financial incentive conditions. Finally, to evaluate H4 (the combined treatment will have the largest joint effect on review volumes and review length), we follow the approach of Burtch et al. (2015) and construct a third outcome measure, unconditional length (populating a length of 0 for those subjects who did not author a review), which jointly captures the combination of quantity and length of reviews, and thereby allows us to evaluate the total influence of our treatments on both outcomes, in tandem.

Beyond testing our main hypothesis in the first experiment, we conduct a variety of analyses to better understand these findings and to triangulate our results. In particular, we replicate our experiment, testing each hypothesis again in Study 2 via a similar combination of treatments. We then explore two additional treatments intended to help identify the specific role of changes in intrinsic motivation, while eliminating

the possible confounding influence of self-selection. In each new experimental condition, we reinforce the social norm or the financial incentive only *after* a subject has agreed to author a review. Under this setup, any differences in review length that might arise could only be attributed to changes in a subject's behavior, conditional on agreeing to participate.

Finally, following the experiments, we further triangulate and clarify our findings in a number of ways. First, we hand-code additional outcome measures from the reviews obtained in Study 1 and Study 2 (namely helpfulness, diagnosticity), which we then analyze to draw a connection between 'review quality' and review length. Second, we collect and analyze a large volume of paid and unpaid online reviews from Amazon.com, via which we demonstrate that financial incentives, in particular, can impact consumers' intrinsic motivation when writing reviews. Finally, we draw on data from both experiments and our sample of archival data from Amazon, to explore possible biases in consumer sentiment that may arise as a result of the treatments.

3.2. Power Analysis

To determine the number of participants needed to have sufficient power to detect our hypothesized effects in our two experiments, we conducted an *a priori* power analysis based on the average effect size obtained in past work most relevant to the current study (Cabral and Li 2015; Goldstein et al. 2008; Mason and Watts 2009; Nolan et al. 2008). The average effect size in this work is a Cohen's *d* of 0.47, which would lead us to require a minimum sample of 73 subjects per condition using two-tailed *t*-tests, with a power of 0.80. If we anticipate a relatively more conservative Cohen's *d* of 0.30, we would require a minimum of 176 participants per condition.

3.3. Study 1

We partnered with a large online retailer located in China that sells children's apparel via TMall, an online platform for business-to-consumer retail. TMall hosts retailers' online sales operations and also allows customers to write and post online reviews about products they purchase. TMall is owned by Alibaba and hosts large businesses, which utilize its marketplace to advertise, promote and sell their products. Businesses on TMall can engage with customers in many ways, such as by offering product promotions, discount coupons, and even issuing targeted SMS text messages. After a purchase transaction,

the buyer can optionally choose to submit a review for the product.[4] Although SMS has traditionally been used by TMall's online retailers to communicate product delivery notices, buyers may also receive promotional SMSs from time to time).[5]

3.3.1. Experiment Design and Procedure

Table 1
Study 1: Treatment Conditions

Condition	Description
No Message	No SMS was issued.
Control	"Dear Buyer, you purchased <<product>> from our online store on <<date>>. Please write a review for this product."
Money	"Dear Buyer, you purchased <<product>> from our online store on <<date>>. Please write a review for this product. You will receive a ¥10 coupon for use in our online store for your effort."
Social	"Dear Buyer, you purchased <<product>> from our online store on <<date>>. Last month, 3,786 other buyers have submitted product reviews for our store! Please write a review for this product.
Money + Social	"Dear Buyer, you purchased <<product>> from our online store on <<date>>. 3,786 other buyers have submitted product reviews for our store last month! Please write a review for this product. You will receive a ¥10 coupon for use in our online store for your effort."

Participants. Our participants included 2,000 customers of our retail partner, well above the threshold specified by our power analysis. Each participant was entered into the sample sequentially over a 2-day period. With each sequential transaction, the associated customer was entered into our sample and randomly assigned to one of five conditions: no message, control, money, social, and money + social.[6]

[4] The merchant and product review systems are owned and maintained by TMall, not the retailers. With respect to valence, only aggregate ratings are displayed on the TMall website and provided to the individual retailers; that is, the valence of individual customer ratings is not observed. However, the platform does provide retailers (and us by extension) with the review text. We therefore initially focus on the textual content in the analysis of our first study.

[5] Using SMS messaging to communicate with customers has many advantages over email. Because cellular numbers are recorded as part of a buyer's shipping information (in China it is common practice for the carrier to contact the buyer via phone call or SMS before delivering an item), a business can maintain greater confidence that communications have indeed been received by the customer. Moreover, it has been reported that as much as 20% of all promotional emails are flagged as spam by email service providers, and thus never delivered to customers.[5]

[6] Customers were excluded from consideration if they had already entered the sample as a result of a prior transaction. Because of our randomization procedure, the offer of a financial incentive and the dissemination of a social norm are independent of other factors that might influence reviewing behavior. To ensure that this was the case, we performed balance tests across a number of available subject-level covariates that we obtained from the retail partner and a third-party market research firm that tracks data about TMall users. Table A1 in the appendix

Our five treatment conditions are summarized in Table 1. For reference, here we provide an English translation of the SMS message, which were confirmed by three coders, fluent in both languages.

Experimental Manipulations. In our no message condition, subjects were not contacted at all. This condition served as a baseline. In our control condition, subjects received a generic SMS message, asking that they write a review for their recent product purchase ("Please write a review for this product"). In our money condition, subjects were asked to write a review using the same language as in the control, and were also told that they would be paid ¥10 upon doing so (approximately $1.50 USD).[7]

When making the review request and offering financial compensation, we did not impose any conditions on the content or quality of the consumer's review (e.g., a minimum length). The decision to not enforce a minimum review length was made to ensure that the design of our financial incentive closely matched with the most commonly implemented financial payment schemes that are used by retailers and review aggregation sites today. For example, AMZ Review Trader, which enables Amazon retailers to offer consumers a product discount in exchange for committing to provide a review upon receipt of the product, encourages lengthy, thoughtful reviews, but does not impose any conditions on compensation beyond the mere posting of a review. Many other major reviewing platforms do not enforce minimum review lengths, including Amazon and TMall, the setting for our experiment.

In our social norm condition, participants were asked to write a review, and were informed that 3,786 different customers had written a product review for the retailer in the prior month (a true value reflecting actual reviewing volumes in the 30 days prior to our experiment). Finally, in our combined money + social condition, participants were asked to write a review, were told that they would be compensated ¥10 upon doing so, and were informed of the number of other recent customer reviewing volumes. When implementing our treatments, we used a "push" approach, delivering communications to participants via SMS messaging. We did not ask buyers to offer a good or positive review; we simply asked that they provide feedback. In addition, we considered the potential for interference to manifest in our experiment (e.g., communication or interaction between subjects in different treatment groups). We examined the geographic distribution of consumers across districts (the Chinese equivalent of a zip code), and observed

---

reports the results of these balance tests. The general lack of significant differences supports the validity of our randomization procedure

[7] This payment amount is in line with past work in this space, which has typically offered subjects $1 to $2 USD in exchange for authoring an online review (Cabral and Li 2015; Stephen et al. 2012; Wang et al. 2012). We also explored an additional treatment condition, in which we offered subjects ¥5. We observed no statistically significant difference between the ¥5 and ¥10 groups for any of our dependent variables.

that our 2,000 subjects were spread across 883 different districts, a wide geographic area. As such, interference is of little concern.

Dependent Variables. Two weeks after the treatment began, the retailer supplied us with information about which customers ultimately authored a review for their product purchase, and the textual content of each review. This allowed us to construct measures of a consumer's effort intensity and the resultant review quality. We examined three primary outcomes in our analyses: (1) participation in review authorship, measured as the quantity of reviews authored, (2) intensity of review authorship, measured as the length of the reviews, and (3) a combined measure, the product of the two, intended to capture the joint impact on quantity and quality.

The logic underlying the third, combined measure is based on recent experimental work that has dealt with a similar two-stage decision-making process. Burtch et al. (2015) considered the effect of a randomized treatment on conversion and contribution amongst visitors to online crowdfunding campaigns. In that context, contribution amounts can only be observed if conversion takes place. This is similar to our setup, wherein review length can only be observed if authorship takes place. Burtch et al. (2015) evaluated the net effect of their treatment on the unconditional expected contribution (i.e., contribution per campaign visitor) by taking the product of binary conversion and continuous contribution amount. The analog in our context is to equate the lack of a review to a review with no content, i.e., a review of 0 length (or minimal helpfulness and diagnosticity, in the case of our quality measures, which we discuss below). Although we recognize that providing a review valence without text is of strictly greater value than not providing a review at all, we make this simplifying assumption for the purposes of assessing the overall benefit of each treatment, as manifest in the resulting overall body of review content.

Additional Measures. In addition to evaluating our main hypotheses on the three key outcome measures, we explore the relationship between proxies for review quality and review length. We operationalize quality in terms of content coded helpfulness and diagnosticity. Based on the literature, a review may be viewed as having high diagnosticity if it helps consumers to identify product attributes, and to characterize those attributes as being either positive or negative (Jiang and Benbasat 2007). In contrast, perceived helpfulness is a more subjective measure, reflecting a buyer's evaluation of how useful a particular review is in coming to a purchasing decision. These dimensions were manually coded for each review by two research assistants, reporting Likert scale values in each case, ranging from 1 to 7, labeled extremely unhelpful (undiagnostic) and extremely helpful (diagnostic) at the endpoints.

To ensure consistent coding, we conducted two instructional sessions, using 35 reviews of products sold by the same merchant (note that these 35 reviews did not come from our experimental sample). In the first instructional session, the concept of review diagnosticity was explained to the coders. The coding assistants and one of the study authors then proceeded to code 10 reviews together, to help the coders better understand the task. In the second instructional session, the students were asked to code the remaining 25 reviews, and to then reconvene, to compare and discuss any coding discrepancies. Following the two instructional sessions, the coding assistants were asked to independently code all of the reviews generated in our experiment, in terms of review diagnosticity and perceived helpfulness. The coders were blind to condition, meaning that they did not know which review came from which experimental condition. We assessed measurement validity and consistency of the coding process via Cronbach's Alpha and Krippendorff's Alpha. Constructing our composite measure of perceived helpfulness from the results reported by our two coders, we observe a Cronbach's Alpha of 0.851 and a Krippendorff's Alpha of 0.706. For our composite measure of review diagnosticity, we observe a Cronbach's Alpha of 0.884, and a Krippendorff's Alpha of 0.781. These values are well in excess of standard cutoffs for acceptable use in the literature (Kline 2000). Additional details of the coding procedure, and the coding instructions, are provided in Appendix B.

3.3.2. Experiment Findings

Our empirical analysis begins with a model-free consideration of any average differences in our three dependent measures: the volume of reviews, the length of reviews, and the combination of volume and length. We then consider the relationship between review helpfulness, diagnosticity, our treatments and review length. Table 2 presents our descriptive statistics for the variables that enter into our analysis. We first consider the impact of each treatment on the probability that a subject authors an online review. The *No Message* group attracted 16 reviews, the *Control* group attracted 27 reviews, the *Money* group drew 73 reviews, the *Social Norm* group drew 41 reviews and the *Money + Social Norm* group drew 70 reviews.

Figures 1a-1c graphically depict the differences across conditions in average review volumes (Figure 1a) and review length (Figure 1b – note that our y-axis reflects the number of Chinese characters; 1 Chinese character translates to roughly 1 English word). We tested H1 and H2 using pairwise comparisons of group means. H1 predicted that offering financial incentives should lead to an increase in the volume of reviews that are provided compared to simply asking consumer to provide reviews. Consistent with H1, we observe that participants in the financial incentives condition were more likely to author a review than the control condition ($p < 0.001$).

Table 2
Study 1: Descriptive Statistics

Variable	Mean	St. Dev.	Min	Max	N
Authorship	0.114	0.317	0.000	1.000	2,000
Length	13.410	8.051	1.000	42.000	227[x]
Log(Length)	2.376	0.730	0.000	3.738	227[x]
Perceived Helpfulness	2.901	1.461	1.000	6.500	227[x]
Review Diagnosticity	2.892	1.362	1.000	7.000	227[x]

Notes: x – 227 subjects wrote a review, out of 2,000 – this value reflects only authored reviews

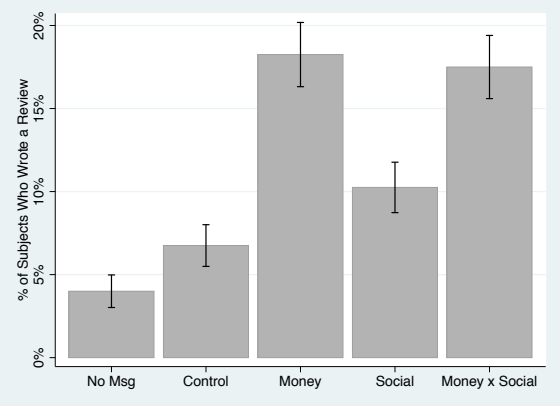

Figure 1a.
Study 1: Percent of Subjects Who Wrote a Review in Each Condition

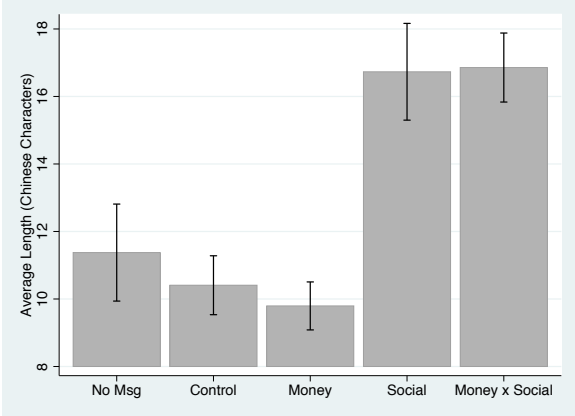

Figure 1b.
Study 1: Average Length of Reviews Written in Each Condition (# Chinese Characters)

We next tested H2, which predicted that providing social norms should lead to an increase in the volume of reviews that are provided compared to simply asking consumer to provide reviews. We found only marginal support for H2, that participants in the social norms condition were more likely to author a review than the control condition ($p = 0.076$). In addition to testing H1 and H2, we also observed that participants in the social norms group were also more likely to author a review than in the no message group ($p < 0.001$), that the money group was more likely to author a review than the social norms group ($p = 0.001$), and that the money and money + social norms groups appear roughly equivalent in their likelihood of authoring a review ($p = 0.782$).

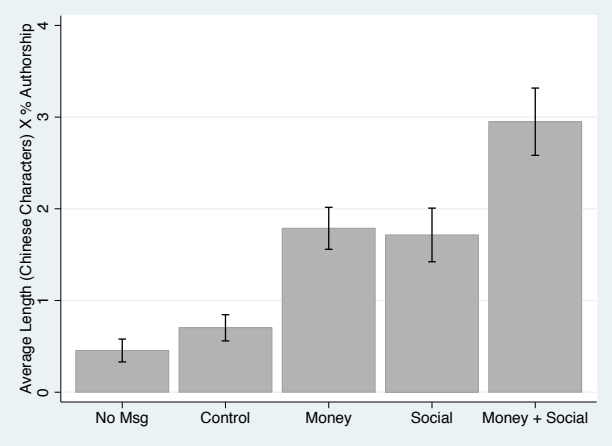

Figure 1c.
Study 1: The Combined Effect on Authorship and Length
Across Conditions (i.e., Length = 0 if No Review Was Written)

Next, we consider the average length of reviews to test H3, which stated that providing social norms should lead to an increase in length of online reviews compared to providing financial incentives or simply asking consumers to write a review. Consistent with H3, participants in the social norms condition wrote reviews that were longer than those in control condition ($p = 0.001$) or those in the financial incentives condition ($p < 0.001$). In addition to testing H3, we also observed that the length of reviews in the control condition were roughly similar to those in the no message group ($p = 0.544$) and the money group ($p = 0.634$). Moreover, the length of reviews in the money + social group authors was longer than in the control group ($p < 0.001$), but there was no discernible difference in the length of reviews in the social norms condition and the money + social condition ($p = 0.942$).

Finally, we test H4 by considering the net effect of our treatments on volume and length, which stated providing social norms and financial incentives, together, should lead to the greatest combination of volume and length of reviews. To measure the 'net' effect, we constructed a combined measure, unconditional length, where we assign a 0 value to length whenever a review was not authored (Figure 3c). Consistent with H4, we find that the money + social group drives total review output in excess of the social norm group ($p = 0.009$), the money group ($p = 0.007$), and the control group ($p < 0.001$). Thus, we find support for H4. In addition to testing H4, we observed that the social norm and the money groups each had a greater combined effect than the control group ($p = 0.001$ and $p = 0.002$, respectively), and that social norm and the money were similar in their combined effect ($p = 0.845$). Finally, we observe that the combined output of the control group and the no message group were also similar ($p = 0.192$). Here,

we should note that if we were to consider conventional thresholds for statistical significance, e.g., an α of 0.05, applying a very conservative Bonferroni correction to account for multiple comparisons would have little impact on our hypothesis tests; H1, H3 and H4 would continue to be strongly supported, while support for H2 would be weak, at best.

To obtain more efficient estimates of the treatment effect, we next consider econometric estimations. We begin by estimating the relationship between our various treatment conditions and the probability of a subject authoring an online review. To conduct this analysis, we relate our binary outcome (review) to dummy indicators of each of our treatment conditions (Equation 1). Subsequently, we employ ordinary least squares (OLS) to analyze the relationship between review length and treatment (Equation 2). As noted above, we also report subsequent regression analyses of our review quality measures (helpfulness and diagnosticity) on our treatments and our effort measure, length, in an attempt to validate the connection between downstream benefits to other consumers that result from increased rates of reviewing and greater intensity of effort exerted by reviewers (Equations 3a, 3b). Here, subjects are indexed by *i*, and our various treatments are indexed by ρ.

$$Authorship_i = \alpha + \sum_\rho Treatment_i^\rho + \varepsilon_i \quad (1)$$

$$Log(Length_i) = \alpha + \sum_\rho Treatment_i^\rho + \varepsilon_i \quad (2)$$

$$Helpfulness_i = \alpha + \sum_\rho Treatment_i^\rho + Log(Length_i) + \varepsilon_i \quad (3a)$$

$$Diagnosticity_i = \alpha + \sum_\rho Treatment_i^\rho + Log(Length_i) + \varepsilon_i \quad (3b)$$

The results of our authorship and length regressions align with the findings reported in our model-free descriptive analyses and graphical depictions with pairwise comparisons of group means (Table 3, columns 1 and 2). In this regression analysis, we also examine the relationship between our treatments, review length and our measures of review helpfulness and diagnosticity. We first estimate a pair of ordinal logistic regressions, taking our coded Likert-scale measures of helpfulness and diagnosticity as dependent variables (columns 4 and 5). We then repeat the process in an unconditional manner, assigning values of 1 to helpfulness and diagnosticity in those cases where no review was supplied (columns 6 and 7). In each case, we find that quality is primarily associated with review length. In summary, Study 1 found support for each of our four hypotheses. First, people were more likely to write a review when they were promised either a small payment or when they were informed of the social norm. The regression results indicate once again that financial incentives (H1) can be very effective in stimulating larger

Table 3.
Regression Results (Study 1)

Explanatory Variable	Authorship	Conditional Log(Length)	Unconditional Log(Length)	Conditional Helpfulness	Conditional Diagnosticity	Unconditional Helpfulness	Unconditional Diagnosticity
No Message	**-0.028+ (0.016)**	-0.009 (0.201)	**-0.064+ (0.038)**	0.265 (0.605)	0.398 (0.585)	0.238 (0.602)	0.308 (0.556)
Money (M)	**0.115*** (0.023)**	-0.184 (0.122)	**0.245*** (0.054)**	0.326 (0.368)	0.316 (0.356)	0.320 (0.371)	0.349 (0.363)
Social (S)	**0.035+ (0.020)**	**0.389** (0.137)**	**0.120* (0.052)**	0.056 (0.407)	**0.752+ (0.448)**	-0.016 (0.410)	0.598 (0.455)
M + S	**0.108*** (0.023)**	**0.382** (0.123)**	**0.317*** (0.062)**	0.298 (0.356)	0.488 (0.367)	0.233 (0.359)	0.357 (0.376)
Log(Length)	--	--	--	**4.135*** (0.355)**	**3.063*** (0.273)**	**4.717*** (0.311)**	**3.921*** (0.187)**
Constant	**0.068*** (0.010)**	**2.248*** (0.087)**	**0.159*** (0.030)**	--	--	--	--
Observations	2,000	227	2,000	227	227	2,000	2,000
F-stat	**18.03 (4, 1995)**	**7.73 (4, 222)**	**18.00 (4, 1995)**	--	--	--	--
Wald Chi^2	--	--	--	**151.50 (5)**	**143.45 (5)**	**243.06 (5)**	**466.16 (5)**
R-squared	0.032	0.126	0.031	0.255	0.205	0.628	0.613

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; + $p < 0.10$; Robust standard errors in parentheses; Regressions based on raw length exhibit the same pattern of results in terms of significance and magnitude.

volumes of reviews, while social norms (H2*) may* be effective. It should be kept in mind that, although we observe that financial incentives were more effective than social norms at eliciting review volumes in our particular experiment, we would be hesitant to conclude that financial incentives are generally more effective because the relative strength of each treatment will depend on the amount of money offered, or the strength of the social norm. Second, people wrote longer and more useful reviews when they were informed of the social norm compared to when they were promised a small payment or received a generic request to author a review. This finding suggests that social norms are effective at stimulating lengthier reviews (H3).

Finally, providing a combination of financial incentives and social norms was most effective at motivating people to write high volumes of lengthy online reviews. This last finding suggests that combining financial incentives and social norms delivers the greatest overall benefit because it jointly stimulates greater review volumes and review lengths (H4).

3.4. Study 2

The second experiment sought to replicate the findings of Study 1 in a different context (to evaluate generalizability across cultures and to a non-purchase setting). Study 2 was also designed to explore the mechanism underlying the observed effects (e.g., self-selection vs. changes in behavior). To do so, we incorporated additional treatment conditions into Study 2, wherein we initially supply the combined money and social norm treatment, but we then reinforce either the financial incentive *or* the social norm after the person has agreed to supply feedback. By comparing the relative efficacy of these two new reinforcement conditions, wherein subjects are retreated after selection has already taken place, we can explore the degree to which the effects we observed are driven by changes in individuals' level of intrinsic motivation versus self-selection.

Were we to observe no significant differences between the two reinforcement conditions, or between those conditions and the baseline, we might conclude that our main effects are driven primarily by individuals selecting into each treatment, who are predisposed toward exerting greater or lesser effort, and thus authoring lengthier or shorter reviews. Conversely, if we were to observe that reinforcement of the social norm does result in lengthier reviews, this suggests that individuals' behavior is modified by the treatment. In each of the new 'reinforcement' conditions, subjects were exposed to a second reinforcement message immediately after they agreed to author feedback.

The reinforcement message reminded people either of the financial benefit of writing the review (e.g., "You will receive $0.04 upon completion of this survey") *or* it reminded them of the social norm (e.g., "You are now the 257th Turker to provide us with feedback"). Thus, all people in the two new conditions were first provided with the combination of financial incentives and social norms when they were choosing whether to write a review; however, after making the decision to write the review, people were subsequently reminded of either the social norm or the financial incentive. We expected that the two new conditions (combined + social norm reminder, and combined + financial incentive reminder) would produce a similar effect on review volumes as the original combined condition, without any reminder, given that the reinforcement message would not be delivered until after a subject had agreed to write a

review. However, we explored whether the two new conditions might have a different effect on review lengths. In particular, we examined whether reminding people of the financial reward could undermine the intensity of effort, leading shorter reviews. At the same time, we also considered whether reinforcing the social norm might have a different effect, e.g., potentially increasing review lengths over and above the baseline combined condition.

3.4.1. Experiment Design and Procedure

In Study 2, we recruited 1,200 Turkers to respond to a survey about the demographics and work behavior of workers on AMT (Ipeirotis 2010). Additional screening questions were embedded in the survey to identify subjects who were not paying sufficient attention (e.g., "what month is it?"). This resulted in the exclusion of 7 subjects. Following a 36-hour delay after completion of the survey, we invited the subjects (via email, using the AMT Requester API) to provide an overall rating of the quality of the demographics survey, in the form of a 7-point scale response, as well as any comments, suggestions or feedback (text).

Subjects were randomized into one of 6 groups: control, money, social, money + social, ***money + social with money reinforced***, and ***money + social with social reinforced*** (note: the bolded treatments are the new treatment conditions that we introduced, to evaluate the mechanism underlying the combined treatment effect and to assess whether the treatment effects operate via changes in intrinsic motivation or self-selection. The first four groups were equivalent to those employed in Study 1. We maintained only 1 control group in this instance because organic (unprompted) feedback was not possible; some form of survey invitation was required. The presence of a financial incentive or a social norm was communicated in the email itself, in both the subject line and the body. Table A2 in the appendix reports randomization (balance) tests for a set of self-reported, subject-level demographic covariates obtained from the initial recruitment (demographic) survey, where we observe no significant differences.

If a subject was assigned to the money group, the email subject line read "Receive $0.04 for providing feedback about our survey!" In contrast, if a subject was assigned to the social norm group, the subject line read "Join the 256 Turkers who have provided us with feedback about our survey!" In the combined conditions, the email subject line mentioned both treatments – i.e., "Receive $0.04 and join the 256 Turkers who have provided us with feedback about our survey!" In the various financial incentive conditions, the email body contained a hyperlink to AMT, pre-populated with task search parameters, thereby directly navigating the subject to group-specific Human Intelligence Task (HIT) on AMT, which in turn linked to a follow-up survey. Subjects were assigned a unique qualification on AMT prior to

running the experiment, to ensure no other individuals could stumble upon the HIT or survey by accident. In the unpaid conditions, the email body contained a hyperlink directly to the follow-up survey.

In each of the new treatment condition, subjects received an additional message upon navigating to the follow-up survey, reinforcing either the financial incentive (e.g., "You will receive $0.04 upon completion of this survey"), or the social norm (e.g., "You are now the 257th Turker to provide us with feedback"). Figures 2a and 2b provide screenshots of the reinforcement treatments. Table 4 provides a summary of our treatment conditions.

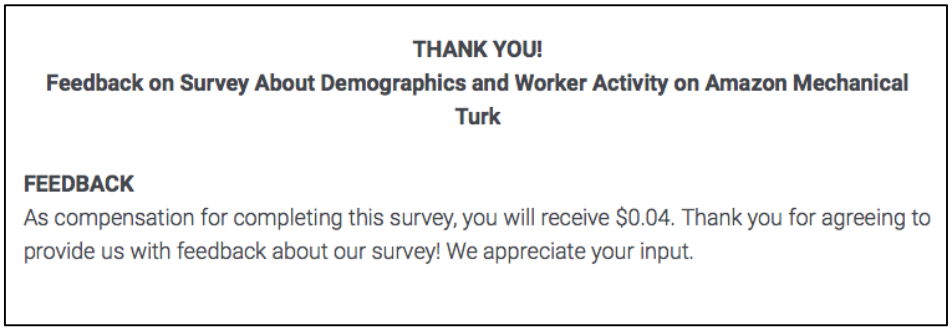

Figure 2a. Reinforcement of Financial Incentive

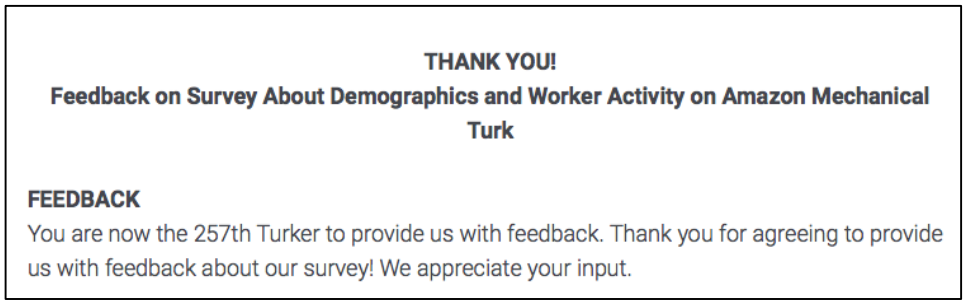

Figure 2b. Reinforcement of Social Norm

Follow-up survey responses were collected over the next 24-hours. We again measured the volume of subjects in each group who provided a follow-up response, and the textual length of feedback each subject provided. Moreover, we once again hand-coded the perceived helpfulness and diagnosticity of textual feedback. Finally, we again consider unconditional measures of length, helpfulness and diagnosticity as well, similar to Study 1, substituting a length of 0, a helpfulness of 1 and a diagnosticity of 1 when subjects did not provide any feedback. We defined diagnosticity and helpfulness in a manner analogous to Study 1 (coding details are provided in Appendix B).

Table 4
Study 2: Treatment Conditions

Condition	Description
Control	A generic email was issued soliciting survey feedback.
Money	An email was issued soliciting feedback, offering $0.04 as compensation, in the form of an AMT task.
Social	An email was issued soliciting feedback, noting that 256 other Turkers had already provided such feedback.
Money + Social	An email was issued soliciting feedback, offering $0.04 as compensation, in the form of an AMT task, and noting that 256 other Turkers had already provided such feedback.
Money + Social + Money Reinforced	An email was issued soliciting feedback, offering $0.04 as compensation, in the form of an AMT task, and noting that 256 other Turkers had already provided such feedback. Upon arriving at the follow-up survey, the financial incentive was reinforced.
Money + Social + Social Reinforced	An email was issued soliciting feedback, offering $0.04 as compensation, in the form of an AMT task, and noting that 256 other Turkers had already provided such feedback. Upon arriving at the follow-up survey, the social treatment was reinforced.

3.4.2. Experiment Findings

Figures 3a-3c plot group means and standard errors, in terms of proportion of subjects providing feedback, length of text provided, and unconditional length, in Figures 3a, 3b and 3c. The Control condition, in which we simply asked for feedback, attracted 66 responses, the Money condition attracted 95 responses, the Social condition attracted 69 responses, the Money + Social condition attracted 95 responses, the Money + Social + Money condition attracted 88 responses and the Money + Social + Social condition attracted 95 responses. Descriptive statistics for Study 2 sample are presented in Table 5.

As seen in Figure 5a, we again observe that the financial incentive is effective in driving participation, the exertion of at least minimal effort (supporting H1). Whereas textual feedback is supplied approximately 30% of the time in control, in the various paid treatments, it reaches nearly 50% (money *vs.* control: p = 0.003, money + social *vs.* control: $p = 0.002$, money + social + money reinforced *vs.* control: p = 0.023, money + social + social reinforced *vs.* control: $p = 0.003$). However, in contrast to study 1, we observe no discernible differences in the volume of subjects supplying feedback between the control and social norms treatment ($p = 0.832$). Thus, we do not observe support for H2. We once again observe that the presence of the social norm in tandem with the financial incentive does not appear to change feedback volumes relative to offering money by itself (p = 0.962).

Table 5.
Study 2: Descriptive Statistics

Variable	Mean	St. Dev.	Min	Max	N
Authorship	0.425	0.495	0.00	1.00	1,193
Length	67.565	70.492	0.00	776.00	508[x]
Log(Length)	3.838	1.021	0.00	6.655	508[x]
Perceived Helpfulness	2.134	0.990	1.00	6.667	508[x]
Feedback Diagnosticity	2.151	0.947	1.00	7.000	508[x]

Notes: x – 508 subjects wrote a review, out of 1,193 – this value reflects only authored reviews

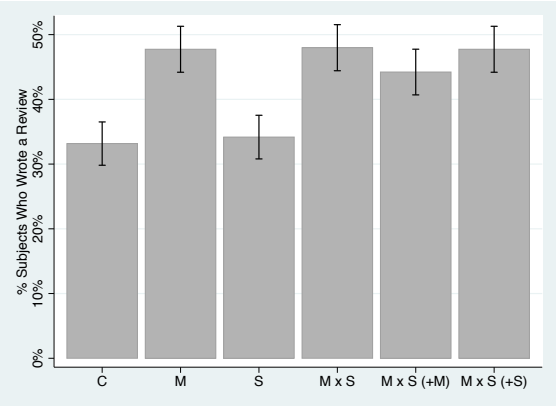

Figure 3a.
Study 2: Percent of Subjects Who Wrote a Review in Each Condition

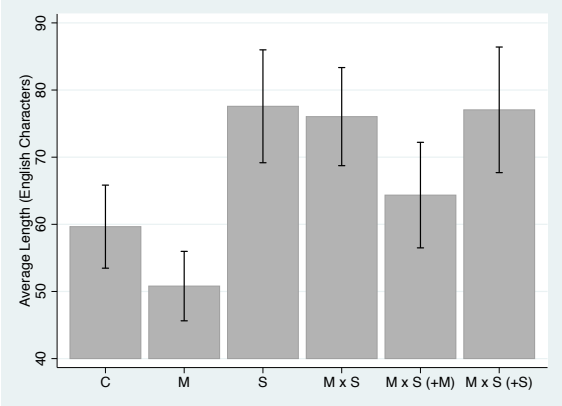

Figure 3b.
Study 2: Average Length of Reviews in Each Condition (English Characters)

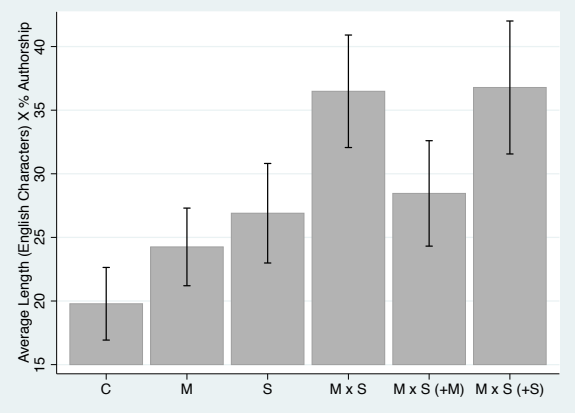

Figure 3c.
Study 2: The Joint Effect on Review Quantity and Length Across Conditions (Length = 0 if No Review Was Written)

Figure 5b reflects a similar pattern to that observed in Study 1 (i.e., Figure 3b). We do not find that simply paying for feedback produces reviews of discernibly greater length (p = 0.274). In contrast, treating subjects with the social norm appears to raise the length of feedback they provide, relative to the common approaches of offering money (p = 0.005) or simply asking, i.e., our control (p = 0.090), once again providing support for H3. Combining social norms and financial incentives, we observe that the effect of social norms on feedback length remains stable; that is, we observe no discernible differences between the social norms group and the money + social norms group (p = 0.891).

Of particular interest, however, are the observed differences between the money + social norm condition, and the two new conditions that involve reinforcement messages. Whereas no stark differences manifest around the number of subjects providing feedback (which is expected, given that reinforcement takes place after the decision to author a review), compared with offering money on its own, offering both money and normative information produced longer feedback than our control (p = 0.005), as did offering the combined treatment along with a reinforcement of the social norm (p = 0.015). However, when the combined treatment was offered along with a reinforcement of the financial incentive, those differences fade (p = 0.146). At the same time, a comparison between the three combined conditions does not indicate discernible differences (money + social vs money + social + money: p = 0.260; money + social vs. money + social + social: p = 0.921; money + social + money vs. money + social + social: p = 0.221).

Finally, considering Figure 5c, which presents group averages for the combined measure (unconditional length), we observe a pattern similar to that observed in Study 1. The combined treatment outperforms the financial incentive (p = 0.023), as well as the social norm condition (p =0.105), once again providing evidence in support of H4. Moreover, when the payment is reinforced, the impact on review lengths from the combined treatment fades, such that it is not discernibly different from the money condition (p = 0.412), whereas reinforcing the social norm causes the increase in effort (length) to persist (p=0.039). Moreover, when we compare the two reinforcement conditions with one another, we do observe relatively clear differences (p = 0.055). At the same time, if we compare the financial incentive reinforcement condition with the baseline combined condition, the differences in review lengths are admittedly less apparent (p = 0.221). Further, the social norm reinforcement condition does not appear to increase review lengths over and above the baseline combined condition, as we might have expected (p = 0.921).

Thus, taken together, the finding in Study 2 offer some evidence in support of our expectations. It appears that the benefits of combining social and financial incentives in this context derive largely from providing subjects with a plausible rationalization for their behavior; it appears that they are choosing to write the

review for reasons other than receiving a financial incentive. This, in turn, results in more intensive provision of feedback (i.e., greater length). At the same time, our results are by no means clear cut, and some between group differences we might have expected to observe ultimately failed to manifest.

We again conducted a formal econometric analysis of these relationships, reported in Table 6. Again, the results of our authorship and length regressions align with the findings reported in our descriptive analyses and graphical depictions with pairwise comparisons of group means (Table 6, columns 1 and 2). Similar to Study 1, we also consider the net effect of our treatment conditions on review output (unconditional length). The results of this regression are reported in column 3. We once again observe positive significant effects in each of the money + social conditions relative to control. Moreover, we find that, compared to the money + social condition ($p < 0.001$), reinforcing financial incentives attenuates the magnitude of the differences in review output, relative to control ($p = 0.018$), whereas reinforcing the social norms does not appear to attenuate the effects ($p < 0.001$). Thus, our regression results are broadly consistent with the graphical comparison of group means noted earlier. However, we also acknowledge that if we make a direct comparison of the coefficients associated with our two reinforcement conditions in the unconditional regression, the differences are rather weak ($F(1, 1187)=1.55$, $p = 0.214$).

Finally, we again assessed the downstream impact of our treatments and length on the helpfulness and diagnosticity of textual feedback. These results are reported in columns 4 and 5. We also considered an unconditional analysis, wherein helpfulness and diagnosticity were coded as values of 1 when no feedback was provided, in columns 6 and 7. Once again, we see that quality measures are primarily driven by our proxy for the intensity of effort that a subject expends, length. One notably difference here is that we observe significant effects on diagnosticity from a number of treatments that involve the social norm. We interpret these results as an indication that effort may manifest in ways other than review length, such as more careful word choice, or greater clarity of writing.

3.5. Study 3

Study 2 provides evidence that our treatments operate at least in part by changing subjects' intrinsic motivation, in that the reinforcement of financial incentives appeared to weaken any benefits of the combined condition. However, that evidence is by no means clear. Moreover, other open questions remain, about the possible collateral effects of our treatments in inducing biased sentiments.

Table 6. Regression Results (Study 2)

Explanatory Variable	Authorship	Conditional Log(Length)	Unconditional Log(Length)	Conditional Helpfulness	Conditional Diagnosticity	Unconditional Helpfulness	Unconditional Diagnosticity
Money (M)	**0.146** (0.049)**	-0.125 (0.193)	**0.473* (0.190)**	0.297 (0.348)	0.547 (0.333)	0.381 (0.346)	**0.652+ (0.334)**
Social (S)	0.010 (0.047)	**0.423* (0.184)**	0.202 (0.194)	0.329 (0.367)	**0.595+ (0.357)**	0.310 (0.364)	0.570 (0.358)
M + S	**0.148** (0.049)**	**0.330+ (0.185)**	**0.699*** (0.200)**	0.373 (0.361)	**0.772* (0.337)**	0.377 (0.362)	**0.768* (0.340)**
M + S (+M)	**0.111* (0.049)**	0.118 (0.184)	**0.456* (0.193)**	0.015 (0.334)	0.482 (0.325)	0.071 (0.335)	**0.562+ (0.328)**
M +S (+S)	**0.146** (0.049)**	**0.373* (0.177)**	**0.710*** (0.199)**	0.291 (0.328)	**0.796* (0.337)**	0.281 (0.329)	**0.800* (0.338)**
Log(Length)	--	--	--	**1.342*** (0.145)**	**1.183*** (0.148)**	**1.645*** (0.086)**	**1.555*** (0.082)**
Constant	**0.332*** (0.033)**	**3.652*** (0.154)**	**1.211*** (0.133)**	--	--	--	--
Observations	1,193	508	1,193	508	508	1,193	1,193
F-stat	**4.16*** (5, 1187)**	**4.06** (5, 502)**	**3.97 (5, 1187)**	**98.96 (6)**	**74.81 (6)**	**393.18 (6)**	**364.08 (6)**
R-squared	0.017	0.040	0.016	0.077	0.071	0.345	0.351

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; + $p < 0.10$; Robust standard errors in parentheses; Regressions based on raw length exhibit the same pattern of results in terms of significance and magnitude.

To gain further clarity on these two issues, we performed a third, archival study, based on a large sample of online reviews collected from Amazon.com, wherein we were able to reliably identify whether or not the retailer had provided a financial incentive to the consumer. Details of the data collection process and sample characteristics are provided in Appendix C. We estimate a linear three-way fixed effects model (product, reviewer and time) on a set of 90,764 reviews for 839 products, authored by 57,469 individuals. We find that the reviews are approximately 6% shorter when a discount was provided by the retailer (column 1 of Table 7). This result is consistent with the findings of Khern-am-nuai and Kannan (2014). More to the point, the result provides clearer evidence of the negative impacts that financial incentives may have on consumers' intrinsic motivation to write reviews. Additionally, this finding further supports our earlier conclusion, from Study 2, that the effects of our treatments are driven, at least in part, by changes in subjects' behavior (e.g., intrinsic motivation), and not just via treatment-induced self-selection.

Table 7.
Regression Results (Amazon Review Trader)

Explanatory Variable	Log(Length)	Positivity
Paid	**-0.06*** (0.012)**	**0.031** (0.010)**
Product Effects	Yes	Yes
Reviewer Effects	Yes	Yes
Year-Month Effects	Yes	Yes
Product-Specific Trends	Yes	Yes
Observations	90,764	90,764
F-stat	**419.62 (730, 57468)**	**145.68 (730, 57468)**
Within R^2	0.101	0.083

Notes: *** $p < 0.001$, ** $p < 0.01$; Robust standard errors in parentheses;

Finally, we consider the question of treatment-induced bias in sentiment, a subject that has received significant attention in the literature, though with inconclusive results. For example, in the experiments by Wang et al. (2012) and Stephen et al. (2012), no apparent differences in review valence were found as a result of payment. In contrast, Khern-am-nuai and Kannan (2014) found that the text of reviews began to contain more positive words, and that the average star valence increased, following BestBuy's introduction of reward-points for writing reviews.

Based on an analysis of hand-coded sentiment in the reviews obtained from our two experiments, we also found no evidence of a treatment-induced sentiment bias. However, we also considered that these null results (and those of prior studies) may have been a result of small sample sizes and thus a lack of study power. Indeed, examining the effect of financial incentives on review valence in our much larger sample of Amazon data, we find that paid reviews are positively biased, being 0.031 stars higher ($p < 0.001$), on average. We report more details of these analyses, along with a more elaborate discussion, in Appendix C.

4. General Discussion

We tested the effectiveness of using financial incentives, social norms, and a combination of both strategies on motivating people to write online reviews. In two randomized experiments, one in the field conducted in partnership with a large online clothing retailer based in China, and a second on Amazon Mechanical Turk, we compared the effectiveness of each strategy at stimulating online reviews in larger numbers and of greater length. We found that financial incentives are more effective at inducing larger volumes of reviews than a simple request (consistent with our first hypothesis, H1), and it appears that social norms may be as well, to some degree (i.e., we observed partial support for our second hypothesis,

H2). When it comes to review length, we found that social norms outperform both financial incentives or a simple request (consistent with our third hypothesis, H3). Finally, we have shown that the combination of financial incentives and social norms yields the greatest overall benefit, motivating reviews in greater numbers and of greater length (consistent with our fourth hypothesis, H4).

4.1. Practical Implications

Many businesses are offering financial incentives to motivate consumers to write reviews. However, using such an approach to solicit reviews appears to present some problems, as discussed earlier. This research suggests that it may be optimal for firms to use financial incentives in tandem with a social norm, in order to minimize the possibility of receiving short reviews. Our results also suggest a more advanced strategy, which may be appropriate for retailers who are launching a new product offering, where little to no reviews have previously been authored by consumers (and thus where advertising a descriptive norm might not be possible). In particular, our findings suggest that firms might initially employ financial incentives to seed early (albeit potentially short) reviews, and then quickly transition to sustainable (lengthier) contributions by exploiting a social norm, ultimately transitioning to higher quality contributions. Of course, this strategy would need to be implemented with caution, because the longer-term effects of social norm treatments remain unclear. In terms of policy, our analyses (particularly those of Amazon.com reviews, reported in Appendix C), suggest that the current policies imposed by federal regulators (e.g., the FTC) and major online platforms (e.g., Amazon) are justified in their view of paid reviews as a possible form of false advertising. Our analyses indicate that paid reviews are systematically more positive than organic reviews, suggesting a bias of reciprocity. Our analyses also suggest that, online retailers may be well served to limit or avoid paying for reviews for reasons beyond simply adhering to regulators and platforms policies, given our finding that paid reviews are systematically shorter in length and, as a result, perhaps of lower quality.

4.2. Theoretical Contributions

While prior research has considered the respective effects of financial incentives (Khern-am-nuai and Kannan 2014; Stephen et al. 2012; Wang et al. 2012) and social norms (Gerber and Rogers 2009; Ferraro and Price 2013; Allcott 2011) in isolation, we offer a first consideration of the relative and joint effects of financial incentives and social norms on motivating pro-social behavior and found that combining financial incentives with social norms result in the greatest overall effect. Further, our work seeks to disentangle the effect of social norms on behavior, distinguishing between the breadth and depth of

engagement in an activity (participation and intensity). Although this distinction has been considered in the financial incentives literature, which has observed in many cases that financial incentives motivate people to do the minimum required to earn pay, no prior work to our knowledge has explored this distinction in the use of social norms to motivate behavior. We observe an asymmetrical effect, in that financial incentives are more effective at motivating participation (write a review) than intensive effort (write a lengthy review), whereas social norms are more effective at motivating intensive effort than they are at motivating participation. Finally, our work contributes to the emerging literature on incentivizing the production of user-generated content (Chen et al. 2010; Jabr et al. 2014; Goes et al. 2016).

4.3. Limitations and Future Research

Our findings around the individual effects of social norms and financial incentives are broadly consistent with the prior literature, in that both appear at least somewhat effective at motivating participation. However, we find that not all participation is equal. Some participants exert less effort than others, whether because of their inherent characteristics or because the treatments we impose cause changes in their behavior. In particular, our three studies we provide evidence that, whereas social norms can drive a high level of effort, as manifest in longer reviews, financial incentives may lead to just the opposite, eliciting shorter reviews. Though the latter finding has not been observed in past experimental work (Stephen et al. 2012; Wang et al. 2012), this difference likely arises because our analysis benefits from a large-scale sample of more than 90,000 reviews. Most interestingly, however, our work is the first to consider and demonstrate that the joint application of financial incentives and social norms can produce the greatest overall benefit, leading jointly to the greatest volume and length of reviews. However, our findings are subject to a number of limitations, which present opportunities for future work.

First, the pattern of effects that we observe across both experiments and the archival study is consistent with the idea that providing financial incentives can undermine intrinsic motivation, placing individuals in an effort-for-payment mindset (Heyman and Ariely 2004). Therefore, we surmise that providing both financial incentives and normative information can circumvent the undermining effect by providing people with a plausible rationalization for their decision to act as one of goodwill, rather than as one of effort for payment. This, in turn, allows intrinsic motivation to persist. However, it is possible (even likely) that our results are also driven in part by subjects' self-selection into the receipt of each treatment. More work is therefore needed to better understand the relative roles of intrinsic motivation and self-selection in driving these outcomes. A replication of the reinforcement conditions from Study 2, with a larger sample, might enable this.

Second, it is also worth noting that we are incapable of comparing the relative effects of financial incentives and social norms on review volumes, because these results are quite likely to depend on the exact level of each treatment (e.g., amount of money offered, the strength of the norm, or how such norms are perceived in different contexts and culture). Future work might explore varied levels of each treatment, to understand how the effects vary. To this point, the effects of financial incentives are quite likely to be non-linear in nature. Cabral and Li (2015) observed that while a $1 rebate had a borderline effect on a consumer's willingness to provide eBay feedback, increasing the amount to $2 produced a stronger effect. Similarly, the effects of social norms are also likely to vary non-linearly in the level of activity amongst an individual's peers. Moreover, our results may also be contingent on the actual level of a subject's own reviewing activity at the time of the experiment. Notably, we have only manipulated the provision of normative information to subjects, we have not manipulated the information itself. This too warrants further exploration.

Finally, it would be useful to understand the dynamic nature of the observed effects, e.g., whether they continue to manifest for a given subject with repeatedly treatments over time, or whether subjects become desensitized. Additionally, it may be possible to improve on our results, by targeting these treatments toward individuals who are most likely to respond in a positive, desirable manner (e.g., can we deliver the financial incentive to only those individuals that exhibit no evidence of a sentiment bias?).

References

Allcott, H. 2011. "Social Norms and Energy Conservation." *Journal of Public Economics* (95:9), pp. 1082-1095.

Anderson, E. 1998. "Customer Satisfaction and Word of Mouth," *Journal of Service Research* (1:1), pp. 5-17.

Aral, S., and Walker, D. 2011. "Creating Social Contagion Through Viral Product Design: A Randomized Trial of Peer Influence in Networks," *Management Science* (57:9), pp. 1623-1639.

Asch, S. 1951. "Effects of Group Pressure Upon the Modification and Distortion of Judgments," In: H. Guetzkow (Ed.), *Groups, Leadership and Men*, pp. 177–190. Pittsburgh: Carnegie Press.

Avery, C., Resnick, P., and Zeckhauser, R. 1999. "The Market for Evaluations," *The American Economic Review* (89:3), pp. 564-584.

Barnes, R. 1949. "Motion and Time Study," New York, Wiley.

Brief, A. P., and Motowidlo, S. J. 1986. "Prosocial Organizational Behaviors," *Academy of Management Review* (11:4), pp. 710-725.

Burtch, G., Ghose, A., and Wattal, S. 2015. "The Hidden Cost of Accommodating Crowdfunder Privacy Preferences: A Randomized Field Experiment," *Management Science* (61:5), pp. 949-962.

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

Cao, Q., Duan, W., and Gan, Q. 2011. "Exploring Determinants of Voting for the "Helpfulness" of Online User Reviews: A Text Mining Approach," *Decision Support Systems* (50:2), pp. 511-521.

Chen, Y., Harper, F., Konstan, J., and Li, S. 2010. "Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens," *American Economic Review* (100:4), pp. 1358–1398.

Cialdini, R. B., Eisenberg, N., Green, B. L., Rhoads, K., and Bator, R. 1998. "Undermining the Undermining Effect of Reward on Sustained Interest," *Journal of Applied Social Psychology* (28:3), pp. 249-263.

Cialdini, R., and Goldstein, N. 2004. "Social Influence: Compliance and Conformity," *Annual Review of Psychology* (55), pp. 591-621.

Cialdini, R., Kallgren, C., and Reno, R. 1991. "A Focus Theory of Normative Conduct," *Advances in Experimental Social Psychology* (24), pp. 201-234.

Cialdini, R. B., and Trost, M. R. 1998. "Social Influence: Social Norms, Conformity and Compliance," In: Gilbert, D., Fiske, S., and Gardner, L. (Eds), *The Handbook of Social Psychology,* Vols. 1 & 2, New York, NY: McGraw-Hill, pp. 151-192.

Deci, E., Koestner, R., and Ryan, R. 1999. "A Meta-Analytic Review of Experiments Examining the Effects of Extrinsic Rewards on Intrinsic Motivation," *Psychological Bulletin* (125:6), pp. 627-668.

Dellarocas, C. 2003. "The Digitization of Word-of-Mouth: Promise and Challenges of Online Feedback," *Management Science* (49:10), pp. 1407-1424.

Dimoka, A., Hong, Y., and Pavlou, P. A. 2012. "On Product Uncertainty in Online Markets: Theory and Evidence," *MIS Quarterly* (36:2), pp. 395 - 426.

Ferraro, P., and Price, M. 2013. "Using Nonpecuniary Strategies to Influence Behavior: Evidence from a Large-scale Field Experiment," *Review of Economics and Statistics* (95:1), pp. 64-73.

Fradkin, A., Grewal, E., Holtz, D., & Pearson, M. 2016. "Bias and Reciprocity in Online Reviews: Evidence From Field Experiments on AirBNB," *Working Paper.*

Frey, B. 1994. "How Intrinsic Motivation is Crowded In and Out," *Rationality and Society* (6), pp. 334-352.

Gallus, J. 2015. "Fostering Voluntary Contributions to a Public Good: A Large-Scale Natural Field Experiment at Wikipedia," *Working Paper. Available at SSRN: http://ssrn.com/abstract=2579118*

Gerber, A.S. and Rogers, T., 2009. "Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You," *The Journal of Politics* (71:1), pp.178-191.

Goes, P. B., Guo, C., and Lin, M. 2016. "Do Incentive Hierarchies Induce User Effort? Evidence from an Online Knowledge Exchange," *Information Systems Research*, Article in Advance.

Goldstein, N., Cialdini, R., and Griskevicius, V. 2008. "A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels," *Journal of Consumer Research* (35:3), pp. 472-482.

Hennessey, B., and Zbikowski, S. 1993. "Immunizing Children Against the Negative Effects of Reward: A Further Examination of Intrinsic Motivation Training Techniques," *Creativity Research Journal* (6), pp. 297-308.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., Cardenas, J., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D, and Ziker, J. 2006. "Costly Punishment Across Human Societies," *Science* (312:5781), pp. 1767-1770.

Heyman, J., and Ariely, D. 2004. "Effort for Payment: A Tale of Two Markets," *Psychological Science* (15:11), pp. 787-793.

Hoffman, M. L. 1981. "Is Altruism Part of Human Nature?" *Journal of Personality and Social Psychology* (40:1), pp. 121-137.

Horton, J., and Chilton, L. 2010. "The Labor Economics of Paid Crowdsourcing," *Proceedings of the 11th ACM Conference on Electronic Commerce,* pp. 209-218.

Ipeirotis, P. 2010. "Demographics of Mechanical Turk," *SSRN Working Paper*. Retrieved from: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1585030

Jabr, W., Mookerjee, R., Tan, Y., and Mookerjee, V. 2014. "Leveraging Philanthropic Behavior for Customer Support: The Case of User Support Forums." *MIS Quarterly* (38:1), pp. 187-208.

Jacques, E., Rice, A., and Hill, J. 1951. "The Social and Psychological Impact of a Change in Method of Wage Payment," *Human Relations* (4), pp. 115-140.

Jenkins, G. D., Mitra, A., Gupta, N., and Shaw, J. 1998. "Are Financial Incentives Related to Performance? A Meta-Analytic Review of Empirical Research," *Journal of Applied Psychology* (83:5), pp. 777-787.

Jiang, Z., & Benbasat, I. (2007). The Effects of Presentation Formats and Task Complexity on Online Consumers' Product Understanding. *MIS Quarterly*, *31*(3), 475–500.

Khern-am-nuai, W., and Kannan, K. 2014. "Extrinsic versus Intrinsic Rewards to Participate in a Crowd Context: An Analysis of a Review Platform." *Working Paper*. Available at SSRN: http://ssrn.com/abstract=2496528

Kline, P. 2000. *The Handbook of Psychological Testing*, (2nd ed.). London: Routledge.

Levi, A., Mokryn, O., Diot, C., and Taft, N. 2012. "Finding a Needle in a Haystack of Reviews: Cold Start Context-Based Hotel Recommender System," *Proceedings of RecSys '12,* Dublin, Ireland, pp. 115-122.

Luca, M. 2016. "User-Generated Content and Social Media," In S.P. Anderson, D. Stromberg, & J. Waldfogel (Eds.), *Handbook of Media Economics* (Vol. 1), pp. 563-592.

Liu, J., Cao, Y., Lin, C., Huang, Y. and Zhou, M. 2007. "Low-quality Product Review Detection in Opinion Summarization," *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 334–342.

Mason, W., and Watts, D. 2009. "Financial Incentives and the 'Performance of Crowds,'" *Proceedings of the ACM SIGKDD Workshop on Human Computation,* ACM.

Mudambi, S., and Schuff, D. 2010. "What Makes a Helpful Online Review? A Study of Customer Reviews on Amazon.com," *MIS Quarterly* (34:1), pp. 185–200.

Nolan, J., Schultz, P., Cialdini, R., Goldstein, N., and Griskevicius, V. 2008. "Normative Social Influence is Underdetected," *Personality and Social Psychology Bulletin* (34:7), pp. 913-923.

Rick, S. I., Cryder, C. E., and Loewenstein, G. 2008. "Tightwads and Spendthrifts," *Journal of Consumer Research* (34:6), pp. 767-782.

Robinson, E., Fleming, A., and Higgs, S. 2014. "Prompting Healthier Eating: Testing the Use of Health and Social Norm Based Messages," *Health Psychology* (33:9), pp. 1057-1064.

Schultz, P., Nolan, J., Cialdini, R., Goldstein, N., and Griskevicius, V. 2007. "The Constructive, Destructive, and Reconstructive Power of Social Norms," *Psychological Science* (18:5), pp. 429-434.

Seuntjens, T. G., Zeelenberg, M., van de Ven, N., and Breugelmans, S. M. 2015. "Dispositional Greed," *Journal of Personality and Social Psychology* (108:6), pp. 917.

Sherif, M. 1936. *The Psychology of Social Norms*, Oxford, England: Harper.

Stephen, A., Bart, Y., Du Plessis, C., and Goncalves, D. 2012. "Does Paying for Online Product Reviews Pay Off? The Effects of Monetary Incentives on Consumers' Product Evaluations," in: *Association for Consumer Research (ACR) Research Conference*. Vancouver, BC: ACR.

Stocks, E. L., Lishner, D. A., and Decker, S. K. 2009. "Altruism or Psychological Escape: Why does Empathy Promote Prosocial Behavior?" *European Journal of Social Psychology* (39:5), pp. 649-665.

Thibaut, J., and Kelley, H. 1959. *The Social Psychology of Groups*, Oxford, England: John Wiley.

Wang, J., Ghose, A., and Ipeirotis, P. 2012. "Bonus, Disclosure, and Choice: What Motivates the Creation of High-Quality Paid Reviews?" *International Conference on Information Systems (ICIS)*, Orlando, FL

Supplementary Appendices:

Appendix A: Randomization Checks

We evaluated pre-treatment balance as a result of our randomization procedure in Study 1 by conducting pairwise Tukey's HSD tests across treatment groups for a number of available subject-level covariates. Our subject pool was randomly assigned across 5 conditions, thus we report 10 pairwise tests for each covariate. The covariates we consider include subject *gender, birth month and day, profile views, reputation* and *mobile purchase.* The latter measure, *mobile purchase*, comes from our retail partner, and is a binary indicator of whether the subject's purchase was made via a mobile device. The remaining four variables were collected, following the experiment, from a third-party market research firm that maintains data on TMall users. We were able to locate the profile data for a subset of our 2,000 subjects (note, we also performed pairwise tests on an indicator of whether the subject was indexed by the third-party market research firm, and found no significance). *Gender* refers to the user gender, coded as 1 for male and 0 for female. *Birth month and day* are respectively coded from 1-12 and 1-31. Because birth month and day, in particular, are self-reported, and thus often missing in the sample, and because there is no recorded value of birth year (preventing a calculation of age), we constructed a binary measure, *birthday*, reflecting whether the user had chosen to report this information or not. *Profile views* is a measure of the number of unique visitors to the users' profile page. Finally, *reputation* is based on a users' various activities on TMall. Accordingly to TMall[9], this value proxies for the reputation of the user on the platform and is based on sellers' feedback about the users in their past purchase transactions.

We observe no notable differences in means for any of the pairwise tests, implying that our treatment groups are balanced in terms of these covariates. In addition to the above, we also have an indication of the product that each subject purchased (item of clothing), as well as their district of residence (akin to a zip code in the United States). Purchases pertained to 18 separate products, and subjects resided in 883 districts. Repeating each of our estimations from Study 1 while controlling for product and geographic fixed effects resulted in no notable differences in our estimates of the treatment effects (in terms of sign, significance or magnitude).

[9] https://service.taobao.com/support/knowledge-847752.htm

Table A1.
Tukey's Pairwise Tests of Pre-Treatment Balance for Study 1

Pairwise Test	Gender	Birthday	Profile Views	Reputation	Mobile Purchase
No Msg. vs. Control	0.5764	3.5721	1.0002	0.2894	0.2006
No Msg. vs. Money	0.7017	0.1282	0.0156	1.5077	0.5041
No Msg. vs. Social	0.3001	0.7061	0.6533	1.2942	0.5041
No Msg. vs. Money + Social	1.6484	0.4691	2.0293	1.1070	2.7076
Control vs. Money	1.2781	3.4439	0.9846	1.7972	0.7020
Control vs. Social	0.8765	2.8660	0.3470	1.5836	0.3008
Control vs. Money + Social	1.0720	3.1030	1.0291	0.8176	2.5071
Money vs. Social	0.4016	0.5778	0.6377	0.2135	1.0028
Money vs. Money + Social	2.3501	0.3409	2.0137	2.6147	3.2090
Social vs. Money + Social	1.9485	0.2369	1.3761	2.4012	2.2062
Observations	1,524	1,524	1,524	1,742	2,000

Notes: Values represent Tukey's HSD pairwise test statistic – critical value for $p < 0.05$ is 3.8623.

We repeated the same process for Study 2, evaluating pre-treatment balance amongst subjects in our sample in terms of all self-reported demographic measures from the initial recruitment survey. The covariates we collected include a variety of factors, including year of birth, to gender, income, education level, country of residence (USA vs India), hours spent on AMT each week, whether the individual is unemployed, to name a few. In Table A2, we report pairwise comparisons using Tukey's test for this set of covariates, noting that no pairwise comparisons were significant for any of the covariates in our survey – these covariates represent a subset of the survey questions detailed in Ipeirotis (2010). Once again, the lack of significance in any of the pairwise tests indicate that our randomization procedure was successful.

Table A2.
Tukey's Pairwise Tests of Pre-Treatment Balance for Study 2

Pairwise Test	Gender	Year of Birth	Income	Education	Country	Hrs / Week	Unemployed
Control *vs.* Social	0.7669	0.8439	0.5977	1.7004	1.7213	0.0356	1.4483
Control *vs.* Money	0.0737	1.1528	0.6886	0.1751	1.4284	0.8188	0.8566
Control *vs.* Money + Social	0.1191	1.6676	0.0675	1.7208	1.1093	0.8360	2.3610
Control *vs.* Money + Social + Money	1.0531	0.0305	2.4265	0.5312	1.4649	1.5633	2.1617
Control *vs.* Money + Social + Social	1.5194	1.0105	2.3268	0.1751	2.2056	0.3210	1.7214
Social *vs.* Money	0.8406	0.3089	0.0909	1.8755	0.2929	0.7832	0.5917
Social *vs.* Money + Social	0.8860	0.8237	0.5302	3.4212	0.6120	0.8004	0.9127
Social *vs.* Money + Social + Money	0.2862	0.8134	1.8288	2.2316	0.2564	1.5277	0.7134
Social *vs.* Money + Social + Social	2.2863	0.1666	1.7291	1.8755	0.4842	0.2854	0.2730
Money *vs.* Money + Social	0.0455	0.5147	0.6211	1.5458	0.3191	0.0172	1.5044
Money *vs.* Money + Social + Money	1.1268	1.1223	1.7379	0.3561	0.0364	0.7445	1.3051
Money *vs.* Money + Social + Social	1.4457	0.1423	1.6382	0.0000	0.7771	0.4978	0.8648
Money + Social *vs.* Money + Social + Money	1.1723	1.6371	2.3590	1.1896	0.3556	0.7273	0.1993
Money + Social *vs.* Money + Social + Social	1.4003	0.6571	2.2593	1.5458	1.0963	0.5150	0.6397
Money + Social + Money *vs.* Money + Social + Social	2.5725	0.9800	0.0997	0.3561	0.7407	1.2423	0.4403
Observations	1,193	1,193	1,193	1,193	1,193	1,193	1,193

Notes: Values represent Tukey's HSD pairwise test statistic – critical value for $p < 0.05$ is 4.037.

Appendix B: Quality Measure Coding Instructions

Study 1 Coding

Two student coders were hired to code the helpfulness and diagnosticity of the reviews obtained in Study 1, from the Chinese-based clothing retailer on TMall. The exact text of the instruction document that was supplied to students tasked with coding the helpfulness of reviews obtained in Study 1, was as follows:

"Helpfulness: code each review on a scale from 1-7, where 1 is least helpful and 7 is most helpful. Suppose you are buying children's apparel for your friend's or relative's child. Please indicate how helpful each review is in terms of enabling you to evaluate the product and make an informed selection.

Diagnosticity: please indicate how many product attributes are mentioned in each review provide. Code each review on a scale from 1-7, using a value of 1 if 0 attributes mentioned, values of 2 through 6 if 1 to 5 attributes are mentioned, respectively, or a value of 7 if 6 or more product attributes are mentioned."

Below, we provide two examples of coded reviews, along with their translations. The first review provides an example of a review that was coded as having low helpfulness and diagnosticity. The second review provides an example of a review that was coded as having high helpfulness and diagnosticity.

- *Low Helpfulness & Diagnosticity Example: "宝贝收到了 很不错 赞一个先" / "Received the item, it is very good, top praise!"*
- *High Helpfulness & Diagnosticity Example: "长度到大腿，红色很好看，洋气，质量也好" / "Length cut to the leg, good looking red color, looks fashionable and has high quality."*

Study 2 Coding

Three student coders were hired to code the helpfulness and diagnosticity of the textual feedback obtained in Study 2, from Amazon Mechanical Turk. In this case, feedback would be viewed as having high diagnosticity if it helps the survey designer to identify survey attributes, and to characterize those attributes as being either positive or negative (Jiang and Benbasat 2007). Examples of this include references to the wording of specific survey questions, the level of compensation offered on AMT for the HIT, the duration of time required to complete the survey, whether questions were intrusive, etc. In contrast, perceived helpfulness, a more subjective measure, reflects a hypothetical survey designer's evaluation of how useful a particular piece of feedback is in revising, improving the survey design. These dimensions were manually coded for each review by three research assistants, once again reporting Likert

scale values, ranging from 1 to 7, labeled extremely unhelpful (undiagnostic) and extremely helpful (diagnostic) at the endpoints.

Once again, to ensure that the coders employed a consistent approach to rating the helpfulness and diagnosticity of each review, we conducted an initial instructional session, using 100 pieces of feedbacks supplied by the Turkers. In the instructional session, the concept of diagnosticity was explained to the coders in the context of this experiment. The coders were told that a survey had been issued to workers on AMT, an online labor market, asking them about demographics and work behavior, and the coders were provided with a copy of the survey. The coders were then instructed to take on the perspective of the survey designer, and assess the degree to which each piece of feedback identified distinct attributes of the survey and whether those attributes were positive or negative (diagnosticity). They were also asked to assess the helpfulness of each piece of feedback, in terms of the degree to which it might help to improve the quality of the survey design.

The students then reconvened with the instructor, to compare and discuss any coding discrepancies. The coding assistants were then asked to independently code all of the remaining feedback generated in our experiment. Once again, the coders were blind to condition, meaning that they did not know which feedback response came from which experimental condition. We once again assessed measurement validity and consistency of the coding process via Cronbach's Alpha and Krippendorff's Alpha. Constructing our composite measure of perceived helpfulness from the results reported by our three coders, we observe a Cronbach's Alpha of 0.921 and a Krippendorff's Alpha of 0.763. For our composite measure of review diagnosticity, we observe a Cronbach's Alpha of 0.951, and a Krippendorff's Alpha of 0.763. Each of these values is once again well in excess of standard cutoffs for acceptable use in the literature (Kline 2000).

The exact text of the instruction document that was supplied to students tasked with coding the helpfulness of reviews obtained in Study 1, was as follows:

"Helpfulness (1 = not helpful at all, 7 = extremely helpful): We have conducted an online survey in which we paid a number of people to report their demographics. The survey is attached for your reference, so you have a sense of the questions that were asked. Following the survey, we asked respondents to provide us with feedback (e.g., suggestions on how to make the questions easier to understand, whether the pay was appropriate or should be adjusted, whether any of the questions felt intrusive, etc.). Please record how helpful each respondent's feedback was in this regard, in terms of how useful or informative the

comments might be for improving the quality of the survey.

Diagnosticity (1 = no elements of survey mentioned, 7 = a wide variety of elements mentioned): How many distinct aspects of the survey were mentioned in the feedback comments (e.g., pay, duration of task, intrusiveness, references to specific questions and so on)?"

Below, we provide two examples of coded textual feedback. The first provides an example of feedback that was coded as having low helpfulness and diagnosticity. The second provides an example of feedback that was coded as having high helpfulness and diagnosticity.

- *Low Helpfulness & Diagnosticity Example: "It was very good I think."*
- *High Helpfulness & Diagnosticity Example: "I thought the questions about age, income and household were well worded and sensible for demographics queries, and the design was clear and easy to navigate. It was just the right length for the pay rate as well."*

Appendix C: Exploring Treatment-Induced Self-Selection and Sentiment Bias

Selection Effects

We collected an archival dataset of online reviews from Amazon.com, for a large sample of products, where we were able to reliably identify whether or not the retailer had provided a financial incentive to the consumer. Amazon's terms and conditions specify that any consumer entering a product review who received compensation for doing so must disclose this fact in the text of the review. To facilitate the identification of such reviews, we first collected data from a third party platform that facilitates the "paid review" process on Amazon for various retailers. This third party platform, AMZ Review Trader, enables Amazon retailers to list product discount offers that consumers can obtain in exchange for committing to review the product on Amazon, following receipt of the product. Moreover, product discounts are the only form of compensation to reviewers that Amazon allows, as long as the discount is disclosed. Because of Federal Trade Commission guidelines (noted in the introduction), AMZ Review Trader, like Amazon, is quite strict in its requirement that consumers disclose the fact that they have received the product at a discount in exchange for providing their review.[10,11] Consumers are required to include a disclosure at the end of their review of the following form: "I received this product at a discount in exchange for my honest and unbiased review." In practice, consumers disclose product discounts using slight variations of this sentence, e.g., leaving out the mention of unbiasedness.

Based on a set of Amazon product listings that were posted on AMZ Review Trader, we identified the unique product identifier, the ASIN, and collected all the associated product reviews from Amazon.com, including the star rating, review text, date stamp, and user ID of the reviewer. We then employed regular expressions to identify the presence of a sentence containing some combination of at least two of the following words: 'discount', 'unbiased' and 'honest'. Upon identifying these sentences, we flagged the review as "paid." We then constructed a measure of review length equal to the number of English characters appearing in the review text, excluding the disclosure sentence. Table C1 provides descriptive statistics for our variables across the 90,764 reviews that were collected. These reviews pertained to 839 products, and were authored by 57,469 individuals.

Equation (4) presents the model that we estimate via a three-way fixed effect specification. Consumers are indexed by *i,* products by *j*, and time (in months) by *t*. Accordingly, the model incorporates consumer,

---

[10] http://help.amzreviewtrader.com/article/122-amz-review-trader-faq#2

[11] http://www.amazon.com/gp/help/customer/display.html?nodeId=201602680: "If you receive a free or discounted product in exchange for your review, you must clearly and conspicuously disclose that fact."

product and time fixed effects, accounting for any unobservable heterogeneity across individuals, products or time periods. This provides us with a plausible basis to rule out self-selection as an explanation for any effect that the *Paid* indicator might have on the length of authored reviews. We present the results of this estimation in the main text of the manuscript, in Table 7 (column 1). We observe that payment does in fact lead to significantly shorter product reviews.

Table C1.
AMZ Review Trader: Descriptive Statistics for Sentiment

Variable	Mean	St. Dev.	Min	Max	N
Valence	4.456	1.052	1.000	5.000	90,764
Length	400.255	450.642	1.000	27,957.000	90,764
Log(Length)	5.451	1.179	0.693	10.238	90,764
Paid	0.399	0.490	0.000	1.000	90,764

$$Log(Length_{ijt}) = \beta * Paid_{ij} + \delta_i + \mu_j + \tau_t + \epsilon_{ijt} \quad (4)$$

Specifically, the reviews that consumers author are approximately 6% shorter when provided in exchange for a product discount. This result is consistent with the findings of Khern-am-nuai and Kannan (2014). More to the point, it further supports our earlier conclusion, from Study 2, that the effects we observe are indeed driven by changes in behavior, and not simply by self-selection into accepting the treatments.

Sentiment Bias

It is also worth considering that our experimental treatments, particularly payment, may have had other, orthogonal effects on the characteristics of online reviews, unrelated to length or quality. For example, paying for reviews may result in inflated star-ratings, if consumers engage in some sort of reciprocity. Of course, this is a chief concern when it comes to paid reviews, both for regulators and other consumers who might read the reviews (Stephen et al. 2012; Avery et al. 1999). Various studies provide mixed evidence on potential for biased sentiment under payment.

For example, Wang et al. (2012) and Stephen et al. (2012) do not observe a significant difference in review valence in response to payment. In contrast, Khern-am-nuai and Kannan (2014) find that the textual content of reviews shifts toward more positive words, and that the average star valence increases, following BestBuy's introduction of reward-points for writing reviews. Finally, Fradkin (2015) observes

a negative shift in average valence of reviews at AirBNB under payment, though he demonstrates that this derives not from a negativity bias as a result of payment, but instead from the fact that non-reviewers on AirBNB are systematically more likely to have had a negative experience. Those individuals do not report feedback because they fear negative reciprocation from the host. This latter finding demonstrates that the potential for shifts in review valence under financial incentives are not straight-forward, and may depend greatly on context.

To assess the possibility of a treatment-induced bias in the valence of online reviews, we conduct a set of additional analyses on the reviews obtained from Study 1 and Study 2. In the case of Study 2, we have a readily available measure by which to assess this bias, because in addition to soliciting textual feedback, we collected a 5-point Likert scale response from subjects about the overall quality of the demographic survey. In the case of Study 1, numerical ratings were not available from TMall at the individual review level. As such, we once again employed (3) student coders to assess the sentiment of the review text. We held a single instructional session in which we briefly explained the task at hand, coding the sentiment of the text of each review from Study 1 as one of negative (-1), neutral (0) or positive (1). The students completed coding of 35 reviews independently, and reconvened with the instructor to ensure shared understanding and agreement on the coding process. The coders then completed the remainder of the task independently. Following completion of the coding process, we evaluated the Cronbach's Alpha and Krippendorf's Alpha of the results, obtaining values of 0.908 and 0.779, respectively. Once again, these values are well in excess of commonly accepted thresholds. Descriptive statistics for sentiment and valence, from Study 1 and Study 2, respectively, are presented in Table C2.

We next regressed the coded ordinal sentiment values from Study 1, and the 7-point scale responses about survey quality obtained in Study 2, on our various treatment indicators, employing Ordinal Logistic regression (Table C3). Doing so, we observe no statistically significant differences from the control group in either case. However, we also observe that the overall models are not statistically significant in either case.

Table C2.
Study 1: Descriptive Statistics for Sentiment

Variable	Mean	St. Dev.	Min	Max	N
Sentiment	0.800	0.480	-1.00	1.00	227[x]
Valence	4.329	0.730	1.00	5.00	508[x]

Notes: x – # of subjects who wrote a review – this value reflects only authored reviews

Moreover, although we observe no statistically significant differences in sentiment or valence under various treatment conditions, we cannot conclude the absence of a bias, because the results may derive from a lack of statistical power. For this reason, we revisit our data on Amazon reviews, which allows us to further assess the possibility of a reciprocity bias deriving from financial incentives. Here, we benefit from a much larger sample of data, and thus a higher statistical power. We employ Ordinary Least Squares regression with product, time and reviewer fixed effects (Table 7, column 2, in the main text). In this case, we do in fact observe a positive, statistically significant effect on star ratings from financial incentives, consistent with the findings of Khern-am-nuai and Kannan (2014).

Table C3.
Regression Results (Studies 1 & 2: Positivity Bias)

Explanatory Variable	**Sentiment**	**Valence**
No Message	0.688 (0.873)	--
Money	0.232 (0.552)	0.459 (0.296)
Social	0.159 (0.596)	0.407 (0.345)
Money + Social	0.336 (0.548)	0.397 (0.291)
Money + Social (+M)	--	0.251 (0.281)
Money + Social (+S)	--	0.397 (0.291)
Observations	227	508
Wald Chi^2	0.78 (4)	3.22 (5)
R-squared	0.002	0.003

Notes: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$; + $p < 0.10$; Robust standard errors in parentheses; Regressions based on raw length exhibit the same pattern of results in terms of significance and magnitude.

Considering recent work which has demonstrated that online product reviews trend from positive to negative over time, we might be concerned that this result is in fact spurious, if paid reviews are more likely to occur early in the product lifecycle. Examining our Amazon data to study the relative timing, within a product, when paid reviews appear, vs unpaid reviews (Figure C1), it does indeed appear that paid reviews tend to appear systematically earlier, in terms of the timing relative to the date the product is initially posted on Amazon's website, and also in terms of review sequence. To address this, we re-estimated the same model, replacing product fixed effects with product-specific time trends, and again using product-specific review sequence trends. In both cases, we observe the same result, a significant,

positive estimated effect of payment, of approximately the same magnitude. As such, this result does not appear to be spurious.

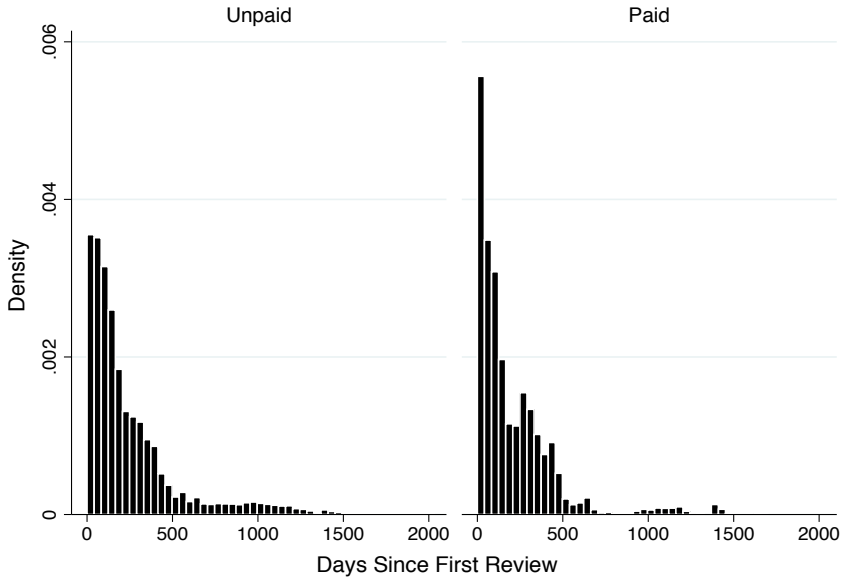

Figure C1. Histogram of Product Age at Time of Review Posting at Amazon.com (Unpaid vs. Paid Reviews)