Schmalenbach Bus Rev (2020) 72:397–435
https://doi.org/10.1007/s41464-020-00094-y

ORIGINAL ARTICLE

The Market for Reviews: Strategic Behavior of Online Product Reviewers with Monetary Incentives

Verena Dorner[1] · Marcus Giamattei[2,3] · Matthias Greiff[4]

Received: 22 October 2019 / Accepted: 1 June 2020 / Published online: 16 June 2020

Abstract Customer reviews reduce search cost and uncertainty about a product's quality. Hence, the quantity and quality of reviews has positive impacts on purchase intentions, sales, and customer satisfaction. In order to increase review quality, retailers and online platforms employ different monetary incentives. We experimentally compare two different incentive schemes: an incentive scheme in which reviewers receive a flat salary, which is independent of review quality, and a tournament incentive scheme in which the reviewer who wrote the most helpful review receives a bonus payment. Helpfulness ratings are assigned by the other reviewers. In our experiment, adverse consequences arise under the tournament incentive scheme. Strategic considerations give rise to strategic downvoting, so that reviewers assign low helpfulness ratings to others' reviews in order to maximize their expected payoffs. Review writing behavior remains unaffected: the tournament incentive scheme does not affect review quality. However, it does destroy the signaling power of helpfulness ratings.

V. Dorner
verena.dorner@kit.edu

M. Giamattei
m.giamattei@berlin.bard.edu

✉ M. Greiff
matthias.greiff@tu-clausthal.de

[1] Institute of Information Systems and Market Engineering, Karlsruhe Institute of Technology, Karlsruhe, Germany

[2] Bard College Berlin, Berlin, Germany

[3] Chair of Economic Theory, University of Passau, Passau, Germany

[4] Chair of Behavioral Management and Economics, Clausthal University of Technology, Clausthal-Zellerfeld, Germany

Keywords Online product reviews · Strategic interaction · Platform markets · Public goods games · Product uncertainty · Tournaments

JEL Codes C91 · D12 · D16 · D47 · M21

1 Introduction

Nearly unlimited choice and pre-purchase uncertainty regarding a product's quality has sparked considerable interest in developing ways of supporting customers' decision making in online shopping (e.g., Brynjolfsson et al. 2003; Gill et al. 2012; Dimoka et al. 2012). To reduce uncertainty, retailers provide manufacturer-independent, customer-written reviews on their websites. Customer reviews are trusted (Bickart and Schindler 2001), increase sales and purchase intentions (Berger et al. 2010; Ghose and Ipeirotis 2011; Park et al. 2007; Chen et al. 2010; for a survey, see Dellarocas 2003), increase post-purchase satisfaction (Stephen et al. 2012), reduce return rates (Sahoo et al. 2018) and increase perceived website usefulness (Kumar and Benbasat 2006). Focusing on the book market, Reimers and Waldfogel (2020) estimate the yearly welfare effects of customer reviews to be $ 41 million for the U.S. Several companies have even built successful businesses around soliciting, collecting and distributing customer reviews to online retailers.[1]

Customers do not, however, consider all reviews but instead search for high-quality and trustworthy information (Chen et al. 2008). Only high-quality reviews will be perceived as helpful and will affect customers' purchasing decisions (Pavlou et al. 2007). Hence, any successful review system will positively affect both the quantity and the quality of reviews. These requirements have led to different incentive schemes, all trying to induce customers to write high-quality reviews. In this paper, we abstract from the quantity-question and analyze how different incentives affect the quality of reviews. More precisely, we compare two different often used incentive schemes: a pay-per-review scheme, in which payment is independent of quality, and a tournament incentive scheme, in which reviewers receive a bonus contingent on the relative quality of the review, as measured by helpfulness ratings assigned by others.

Several retailers run a pay-per-review scheme. Customers receive a fixed payment or a product in exchange for a review. Some review platforms (e.g., www.ilovetoreview.com) connect retailers, who are willing to give away their products for free in exchange for a review, and customers, who agree to write a review in exchange for the product. This could be problematic because, if the costs of writing a review increase with review quality, a money-maximizing customer will write a large number of low-quality reviews.[2]

[1] Some prominent examples are Tripadvisor, Yelp and Foursquare.

[2] There are lots of market research programs where customers can keep the products after testing and reviewing (see e.g., https://www.lifewire.com/programs-to-review-products-and-keep-them-4158347). Although customers do not get a fixed moentary reward, their reward (keeping the product) is in most cases independent of review quality.

To generate high-quality reviews, some retailers condition payment on relative review quality. In order to judge a review's quality, retailers look at the helpfulness votes, which are assigned by customers. In its incentive program "Vine Club", a selected group of Amazon customers receive pre-release products free of charge in exchange for a review. The specific eligibility criteria for Vine reviewers are not publicized, but Amazon admits that the helpfulness of previously written reviews plays a large role. Essentially, "Vine Club" is a tournament incentive scheme in which the bonus consists of being admitted to and remaining in the Vine program. A similar tournament incentive scheme is Yelp's Elite Squad program, in which selected reviewers are invited to exclusive events.[3] A tournament incentive scheme can give rise to strategic downvoting, as the following example illustrates.

Soon after the initiation of Amazon's Vine program, Vine members started to complain that their reviews were accumulating inexplicably high numbers of negative helpfulness ratings. They suspected that fellow reviewers were systematically "voting down" their reviews to oust and replace them as Vine Club members, or to protect their own membership status.[4] Here, an implicit assumption is that a review's quality is measured in relative terms. By assigning a bad helpfulness rating to a fellow's review, the reviewer's own review will be perceived as better relative to the fellow's review.[5] Because the incentive to assign the lowest possible helpfulness rating is strategic (retaining VINE membership), we call such behavior strategic downvoting.[6]

This example reveals a potential weakness of monetary incentives which are based on relative quality, as measured by helpfulness ratings: Due to strategic downvoting, helpfulness ratings may be biased. Possibly, this could discourage reviewers and lead to a decrease in review quantity and quality. It may also imply that a review's helpfulness rating ceases to signal the review's quality to other customers. These problems exist not only if most customers are motivated by quality-based monetary incentives. But, it may also be relevant if there is a large population of silent customers, who never write reviews, as we will argue in Sect. 5.

[3] As with Amazon's VINE program, the criteria for being awarded membership in Yelp's Elite Squad are not publicized. but quantity and quality of reviews plays a large role (see https://www.yelp.com/elite). In an analysis of Yelp reviews, Dai et al. (2018, p. 319) found that elite reviewers do write better reviews, i.e., reviews that are more consistent with peer ratings.

[4] It is not known whether or how Amazon reacted to these accusations. In 2016, Amazon changed its policy. Since 2016, incentivized reviews (i.e., a free product in exchange for a review) are "prohibit[ed] (...) unless they are facilitated through the Amazon Vine program." (https://blog.aboutamazon.com/innovation/update-on-customer-reviews, accessed Nov. 29, 2018). Despite the change in rules, it seems that incentivized reviews still exist (https://www.businessinsider.de/amazon-bad-review-practices-crackdown-2018-4?r=US&IR=T, accessed Nov. 29, 2018). In fact, there is evidence that reviewers actually exchange (fake) positive reviews for monetary compensation (https://www.buzzfeednews.com/article/nicolenguyen/amazon-fake-review-problem#.ewd5B0a8B, accessed Nov. 29, 2018).

[5] See https://annerallen.com/2016/10/amazons-new-review-rules-should-authors-worry/, accessed Nov. 29, 2018.

[6] Even after some recent changes, in which Amazon removed the possibility to rate a review as "unhelpful", there is still a strategic motive when it comes to rating reviews. That is, a reviewer maximizes her chances of remaining a VINE Club member by refraining from marking all reviews written by others as "helpful".

The empirical evidence so far is inconclusive. Closest to our study are Stephen et al. (2012) and Wang et al. (2012). Stephen et al. (2012) show that a flat salary ($ 1 per review) increases the helpfulness of the reviews.[7] Wang et al. (2012) find no effect of a quality-contingent payment ($ 0.25 per helpfulness point) on review helpfulness. A systematic comparison of a flat salary and a tournament incentive scheme has, to the best of our knowledge, not been conducted. Our paper tries to fill this gap.

Using a controlled laboratory experiment and an online survey, we compare a pay-per-review scheme and a tournament incentive scheme. Participants write product reviews and vote about the helpfulness of others' reviews. Writing reviews is implemented as a public good setting where the whole group profits from written reviews. We compare how the different incentive schemes affect review quality and the assignment of helpfulness ratings. The experiment allows us to account for confounding factors such as collusion or social pressure, which are present in the field. Also, in reality, many customers write reviews only when they are extremely happy or frustrated with the product. In the experiment, each participant has to write reviews, which allows us to abstract from the question "Who writes a review?" and allows us to isolate the effect of incentives on review quality. Most importantly, the random assignment to treatments implies that differences in review quality cannot be explained by differences in reviewers' experience or reliability. The exogenous randomization in experimental settings provides a clear methodological advantage over using field data (Falk and Heckman 2009; Bardsley et al. 2010).

The remainder of the paper is organized as follows. We discuss the theoretical background and derive our hypotheses in Sect. 2. In Sect. 3, we introduce the designs of experiment and survey, before we present the results in Sect. 4. In Sect. 5, we discuss our results and the limitations of our research. We conclude in Sect. 5 by deriving implications for managerial practice.

2 Theoretical Background and Hypotheses

2.1 Reviews as a Public Good

Interactions between reviewers share some characteristics to interactions between individuals in a public good game (PGG).[8] Reviews are non-excludable and non-rivalrous. Each potential customer can access the reviews for free. A customer who reads a review does not reduce the benefit others can derive from the review. Hence,

[7] This is in line with the large scale analysis by ReviewMeta (https://reviewmeta.com/blog/analysis-of-7-million-amazon-reviews-customers-who-receive-free-or-discounted-item-much-more-likely-to-write-positive-review/, accessed Nov 29, 2018).

[8] In a typical abstract PGG, each player contributes private resources to a public good, which benefits the group. For each player, the cost of a contribution exceeds her private benefit but the benefit for all group members exceeds her cost. Players maximize their own payoff by contributing zero. However, each group member would receive a higher payoff if every player made a positive contribution. For a game-theoretical description, see Mas-Colell et al. (1995, chapter 11.C), for reviews of the experimental evidence, see Davis and Holt (1993), Ledyard (1995), Zelmer (2003) and Chaudhuri (2011).

reviews constitute a public good, and the quality of the public good increases in the quantity and quality of the reviews.

Writing a review generates private costs and benefits others. Not writing a review and saving these costs is the dominant strategy if a reviewer only cares about her own monetary payoff. If all reviewers follow this strategy, no reviews are written. This constitutes a Nash equilibrium because no individual reviewer can increase her expected payoff by providing higher quality reviews. Clearly, this equilibrium is inefficient, because no information is shared. To realize benefits from customer-written reviews, retailers need to motivate their customers to write as many helpful reviews as possible.

2.2 Approval and Disapproval as Nonmonetary Rewards

When modeling review writing as a public good game, theory predicts that no reviews are written if customers only care about their monetary payoff. Obviously, this is at odds with the large number of reviews that are written. One explanation for the positive number of reviews is that review-writers do not care about money alone but also about nonmonetary rewards. More specifically, if helpfulness ratings can be assigned to reviews, these ratings can be used to express approval or disapproval, which are nonmonetary rewards.[9]

This explanation receives support from the public goods literature. The general pattern in laboratory PGGs is that, on average, participants contribute approximately 50% of their endowment and contributions decrease over time (e.g., Zelmer 2003). Contributions can be increased by adding a second stage, in which participants can allocate nonmonetary rewards or punishment points to each other (Masclet et al. 2003; Dugar 2013; Greiff and Paetzel 2015). Nonmonetary rewards and punishments are usually modelled as the expression of approval and disapproval points. In the literature, this is often referred to as the exchange of social approval, peer approval, or the expression of informal sanctions. In the following, we will refer to it as approval or disapproval. A second theoretical argument for the effectiveness of approval and disapproval applies to repeated games. Approval and disapproval can serve as pre-play communication for future rounds (Masclet et al. 2003, p. 367). Although this form of pre-play communication is cheap talk, it is well-known that cheap talk positively affects contributions (Ledyard 1995; Zelmer 2003).[10] In our experiment, participants can use helpfulness ratings to assign approval and disapproval points

[9] From a theoretical perspective, approval and disapproval can affect contributions if approval and disapproval are arguments in a player's utility function. Masclet et al. (2003) show that in a PGG with homogeneous endowments, approval increases average contributions. Greiff and Paetzel (2015) show that in a PGG with heterogeneous endowments, approval increases average contributions.

[10] In addition, the experiment reported in Gächter and Fehr (1999) reveal that in repeated PGGs with partner matching, the effect of approval is strongest if group identity is established before the game. Dugar (2013) compares the effect of approval and disapproval in a PGG with partner matching. He finds that disapproval points have a larger effect than approval points, but that the effect on contributions is largest when participants are allowed to choose between approval and disapproval. There is also some evidence for antisocial punishment (see Herrmann et al. 2008), which means that high contributors are punished.

about the review written by another participant. This can signal e.g,. that the review is too short and contribution to the public good is insufficient.

Helpfulness ratings provide incentives for review-writing even if helpfulness ratings do not affect the chances of receiving any future monetary reward. However, if good helpfulness ratings increase the chances of receiving a bonus (as in our bonus treatment), this could lead to crowding-out and strategic downvoting, which we discuss next.

2.3 Crowding-Out

A monetary bonus paid for the best review could increase the quality of reviews. However, introducing such a bonus could come with negative psychological side effects, which may weaken the individual's intrinsic motivation to engage in the incentivized activity (e.g., Deci et al. 1999; Gneezy et al. 2011). These effects are referred to as crowding-out. Crowding-out occurs because the strength of the intrinsic motivation depends on the perception of the activity. In the remainder of this section, we will argue that a change in the incentive scheme will affect perception and intrinsic motivation.[11]

Assuming that reviewers are motivated not only by monetary incentives, we can distinguish between extrinsic and intrinsic motivation. In the case of review writing, extrinsic motivation consists of monetary and nonmonetary rewards (i.e., the helpfulness ratings discussed in the preceding section). In contrast to extrinsic motivation, intrinsic motivation derives from rewards inherent to the activity of review writing.

The strength of intrinsic motivation is not independent from monetary incentives. Consider a setting in which a fixed monetary reward is paid for any review, regardless of the review's quality. In this setting, writing a high-quality review is likely to result in feelings of generosity and competence. Feelings of self-interest are unlikely because the monetary reward is independent of quality. By writing a better review, the writer provides more information for others but cannot increase her own payment.

This might be different in a setting in which only the best review is rewarded with a monetary bonus. In this setting, writing a high-quality review is less likely to result in feelings of generosity and competence. This is because review writing might now be perceived as driven by pursuit of the bonus. By writing a better review, the writer provides more information for others but at the same time increases her own expected payment. Hence, when quality is incentivized, a high-quality review is more likely to signal the reviewer's self-interest and can lead to reduced feelings of generosity and competence. Ultimately, this can lead to lower intrinsic motivation as compared to the former setting without a quality-contingent bonus.

[11] For surveys about crowding effects, see Deci et al. (1999), Frey and Jegen (2001), and Gneezy et al. (2011); for a theoretical model, see Benabou and Tirole (2003).

2.4 Strategic Downvoting

Whenever helpfulness ratings are assigned solely based on quality, the best review will receive the highest rating. At first glance, rendering rewards directly dependent on quality appears a straightforward and effective idea (Wang et al. 2012). It rests on the assumption that helpfulness ratings are cast honestly.

However, if only the best review is rewarded with a bonus while all other reviews are unpaid, the setting resembles a winner-takes-all-tournament. In such a tournament the winner is determined by relative quality. Quality is assessed by helpfulness ratings and reviewers have an incentive for strategic downvoting: Assigning a lower rating to others is a form of sabotage (see Harbring and Irlenbusch 2011; or section 6.1 in Dechenaux et al. 2015) and increases the chance of getting the highest rating and receiving the bonus.

Strategic downvoting can reduce reviewers' motivation as well as the signaling power of helpfulness ratings. If helpfulness ratings are a form of approval, as we have argued above, reviewers might expend a lot of time and effort to write a review because this increases the chances that the review will receive a high helpfulness rating. Strategic downvoting might weaken approval because of a lower correlation between helpfulness ratings and quality.[12] Reviewers will learn about or anticipate strategic downvoting, and helpfulness ratings will lose their motivating power. If reviewers anticipate strategic downvoting, the quality of reviews might deteriorate because of crowding-out effects.

A related issue is the effect of strategic downvoting on the signaling power of ratings. If customers expect strategic downvoting, they cannot distinguish between the most helpful reviews by looking at the helpfulness ratings. This means that customers may end up with basing their decisions on mediocre reviews or may not even use the reviews at all.

2.5 Hypotheses

We focus on review systems in which the quality of reviews is endogenously determined by reviewers' helpfulness ratings. Our main goal is to investigate the effect of a quality-contingent bonus on the quality of reviews.

Since reviewers' helpfulness ratings are central to our theoretical reasoning, we start by investigating the effect of a quality-contingent bonus on reviewers' assignment of helpfulness ratings. Based on the theoretical considerations discussed in Sect. 2.4, we derive the following hypotheses, which we will test using data from a controlled laboratory experiment and an online survey. A critical discussion at the end of the paper will further analyze our assumptions behind our hypotheses.

Hypothesis 1 (H1—Strategic Downvoting) Incentivizing review quality by introducing a quality-contingent bonus leads to strategic downvoting.

[12] Several studies on public good games report a strong positive correlation between contributions and approval (see Sect. 2.2). In these studies, however, individuals had no incentive for strategic voting.

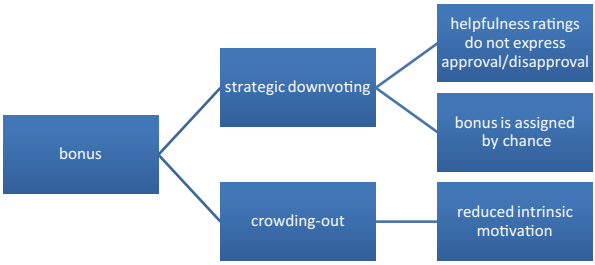

Fig. 1 Theoretical background for Hypothesis 2

The theoretical background for H1 is based on our discussion of strategic downvoting. If quality is incentivized, reviewers maximize their expected payoff by assigning the lowest helpfulness rating to all other reviews. In order to investigate the effect of a quality-contingent bonus on the quality of reviews, we derive our second hypothesis, H2.

Hypothesis 2 (H2—Crowding-Out) Incentivizing review quality by introducing a quality-contingent bonus decreases the average quality of reviews.

The theoretical background for H2 is based on our discussion of crowing-out and strategic downvoting and is illustrated in Fig. 1. If the introduction of a quality-contingent bonus leads to crowding-out, intrinsic motivation will decrease and, consequently, the average quality of reviews will decrease. This holds even if helpfulness ratings are assigned honestly.

In addition to the crowing-out effect, there could be an effect from strategic downvoting. In the extreme case, all reviewers assign the lowest helpfulness rating to each review. This implies that helpfulness ratings are not used to express approval or disapproval, so that the extrinsic incentives from helpfulness ratings are absent. More importantly, strategic downvoting implies that nobody can increase her probability of winning the bonus because helpfulness ratings are statistically independent of review quality, so that the winner is decided by chance alone.[13] Overall, both effects reduce review quality.

3 Experiment and Survey

3.1 Experimental Design

Each participant plays within a group of 5 players for 4 rounds. Each round consists of two stages (see Fig. 2). Group composition remains constant across rounds.

In the first stage of each round, participants receive a product and are given the opportunity to sample it. All participants receive the same product, and, in each round, they receive a new product. Products are shown in Fig. 3.

[13] If strategic downvoting is less extreme, helpfulness ratings will correlate with review quality, and reviewers could increase their expected monetary payoff by increasing review quality. However, the non-monetary incentives from helpfulness ratings will still be lower.

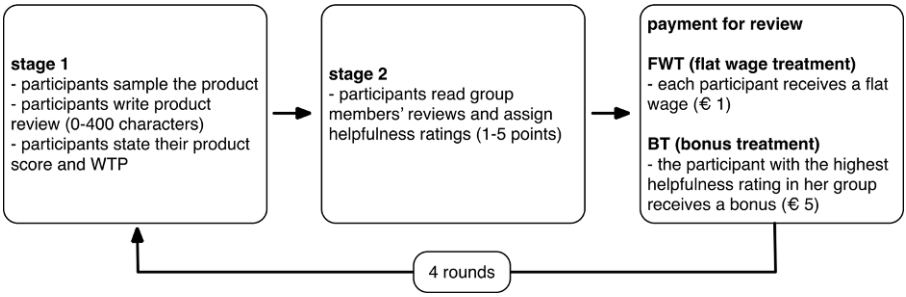

Fig. 2 Experimental design

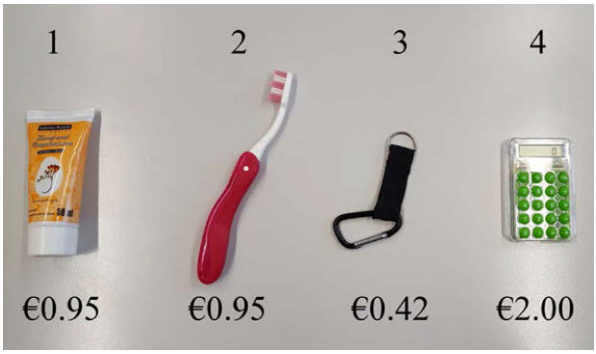

Fig. 3 Sample products in the experiment. (*1* hand balm, *2* tooth brush, *3* key ring, *4* pocket calculator. Prices are displayed below the items)

Products are distributed at the beginning of each round and collected after each round. We use inexpensive everyday experience products in order to (i) minimize the effect of differences in participants' product (type) expertise and (ii) make sure that participants were able to quickly familiarize themselves with the products. Products are always presented in the same order to avoid product order-related confounding. We supplied participants with real products and designed the computer interface similar to existing websites where customers review products. These choices were made so that participants could be expected to be familiar with the environment. Also, instructions were framed to put participants in the shoes of real reviewers.

After sampling the product (without a time limit), participants have to write a product review with 0–400 characters. In addition to the cognitive effort, writing the review is costly in terms of money. For each character written, a participant's payoff is reduced by € 0.004. If a participant writes a review of maximum length (400 characters), € 1.60 are deducted from her endowment. To avoid losses, participants are endowed with € 1.60 per round.

Implementing a fixed cost per character written increases the opportunity cost of review writing. This feature of the design ensures that writing a review is costly, even if costs of real effort (which are unobservable) are close to zero. Moreover, it ensures that participants have no incentive to write uninformative reviews.

Providing reviews generates a benefit for others. We proxy this positive externality by the number of characters. To account for the positive externality, a participant's payoff increases by € 0.005 for each character written by another group member.

Note that participants only profit from reviews written by others. Thus, participant i's payoff from stage 1 is given by:

$$\pi_i = 1.60 - 0.004c_i + 0.005c_{-i},$$

where c_i denotes the number of characters written by participant i and c_{-i} denotes the number of characters written by all other participants $-i$ in the same group. This incentive scheme resembles a public goods dilemma. In the standard PGG, the public good is given by the sum of all participants' contributions multiplied by a positive constant (the so-called marginal per capita return). In our setting, a participant's own contribution does not increase the public good, i.e., it affects only the other participants' payoffs but not the contributor's payoff. Similar to a PGG, the social benefit of a contribution (€ 0.005 for each participant) exceeds the private benefit.

In addition to writing the review, we gathered information about participants' product-specific preferences. Participants had to evaluate the product (numerical product score from 1 = very good to 6 = very bad) and were asked to specify their willingness to pay (WTP). If the WTP exceeded the price of the product (which was unknown to participants), they had to buy the product at this price.[14] The fact that participants did effectively purchase the product (if WTP > price) increases the experiment's realism and external validity. WTP and product score remained private information.

In the second stage, participants were presented with the reviews of all other participants in their group and asked to rate each review's helpfulness on a five-point scale (5 = very helpful, 1 = not helpful at all).[15] So, in our experiment, all participants act both as reviewers and as readers (we discuss this in Sect. 5). We used different scales and input formats for product evaluation in stage 1 and helpfulness rating in stage 2 to avoid confusion among participants.

Using a between-subjects design, we compare behavior across two treatments. Therefore, participants' payoff in stage 2 is treatment specific. In our first treatment, each participant receives a flat salary of € 1 (flat wage treatment FWT)[16]. In the second treatment, the reviewer with the highest average helpfulness rating receives

[14] This procedure is similar though not identical to the Becker-DeGroot-Marschak (BDM) mechanism (Becker et al. 1964). This mechanism ensures that no participant can gain by misrepresenting her true valuation, even if the participant knows the price of the product. In the original BDM mechanism, participants have to buy the product if the WTP exceeds a randomly drawn number. In our variant, participants have to buy the product if the WTP exceeds the price (which lies in a range participants might anticipate). The main difference is the price that participants have to pay in case of purchase. In the original BDM, they would have to pay their WTP; in our variant they have to pay the product's price. We chose this particular variant over the original BDM because a pilot study revealed that participants' WTPs significantly exceeds product prices in many cases. With the original BDM, this would have generated a large income for the experimenters at the expense of our participants.

[15] At the beginning of stage 2, reviews were screened by the experimenters. Nonsensical reviews (e.g. repetitions of the letter "x") could thus have been excluded from the computation of social payoff. Participants were informed about this possibility. But in fact, no such reviews were written.

[16] While many real-world platforms offer free products or vouchers instead of an actual monetary wage we opted for that design choice to have experimental control and keep comparability between treatments. As vouchers may be perceived differently by different participants, we opted for a monetary evaluation. For comparability, the flat wage is equal to the expected payoff in the bonus treatment (chance of getting

a bonus of € 5 while all other group members receive no payment (bonus treatment BT). In case several reviews attain identical helpfulness ratings, the bonus was split evenly between these reviewers. At the end of each round, each participant was informed about her payoff and the average helpfulness rating for her review.

The total payoff of each round is given by the sum of payoffs from each stage. That is, total payoff is composed of (i) the payoff π_i from stage 1 (see the formula above) minus the price of the product if the participant bought the product, and (ii) the payoff from stage 2.

The Nash equilibrium for this game would be to assign the lowest helpfulness rating to every review in BT, while there would be indifference on this decision in FWT. In BT, all reviews would be rated with the lowest helpfulness rating so that the bonus would be split equally among the group of five, yielding the identical payoff as in FWT. Knowing that the length of the review does not influence the chance of getting the bonus would induce participants to write no review in the first stage. As the game is finite, backward induction translates this result from the last round to all previous ones. In this case, participants would earn € 1.60 from stage 1 and € 1 from stage 2, yielding a total payoff of € 2.60 per round (minus the expenditure for buying products).

After the fourth round, participants were asked to complete a questionnaire on demographic variables, product reviewing experience and review usage. The total payoff earnt in the experiment is given by the sum of all rounds' payoffs.

3.2 Survey Design

According to H1, helpfulness ratings, as expressed by participants in treatment BT, may be biased downward due to strategic downvoting. In order to gather unbiased helpfulness ratings, we complemented our experiment by a survey with different participants than in the experiment.

In the survey, participants were asked to rate the helpfulness of reviews written in the experiment. More precisely, we randomly selected the reviews of 20 participants from the experiment (10 from each treatment). The 80 reviews written by these experiment participants (4 reviews written by each participant) were then rated by the survey participants.[17]

3.3 Experimental Procedure

The experiment was conducted at the Passau University Experimental Laboratory (PAULA) using classEx (Giamattei and Lamsbdorff 2019). Upon arrival, participants were randomly seated in the laboratory and given detailed experimental

the bonus is 1/5 * 5 € = 1 €). Different forms of payments (voucher in FWT vs. monetary bonus in BT) would have confounded the results.

[17] Reviews were presented in exactly the same order as they were presented in the experiment. Survey participants used the same 5-star scale as in the experiment. As a difference, participants in the survey, did not sample the products.

instructions.[18] A pre-test and several control questions ensured that participants understood the instructions correctly.[19]

We conducted in 6 sessions with 90 participants[20] in 18 groups (8 FWT; 10 BT). 74% of the participants were female resembling the composition of students in Passau. 13% studied an economics-related major and the average age was 23.3. The experiment lasted on average 91.24 min (sd 16.19). The average payoff of € 14.37 (including the show-up fee of € 2) for around 90 min of work is slightly above an average student salary in Passau at that time.

In the survey, we ensured that the 96 survey participants had not taken part in the experiment. Each participant rated up to 20 reviews.[21] As we recruited students from big lectures, we awarded three gift vouchers with a value of € 10 each. Participants needed 10 min on average to complete the survey. In sum, the survey resulted in 1781 helpfulness ratings, 420 for reviews from FWT, 1361 for reviews from BT. On average, each review from FWT (BT) was rated by about 10 (34) survey participants. We collected more helpfulness ratings for BT as we expected downvoting to be more prominent in this treatment. For our statistical analysis, we will use the average helpfulness ratings from survey participants for the 80 reviews (40 from each treatment, see Sect. 3.2).

4 Results

4.1 Strategic Downvoting and Review Quality

The analysis in this section focuses on three variables: review length as a proxy for quality[22], the helpfulness ratings from the experiment, and the helpfulness ratings from the survey.

[18] Instructions can be found in Appendix A.

[19] We conducted a pre-test with 10 participants each. Both treatments and the questionnaire were tested twice. Pre-test participants were asked to write down suggestions for improvement and requests for clarification of the experimental procedure during the experiment. We implemented the suggestions and found after the second pre-test that all participants had correctly understood the experimental tasks and had not experienced any problems in carrying them out. We also used the pre-tests to calibrate the parameter settings for the maximum number of characters per review, the "exchange rate" for characters and €, and the incentive payment. Participants were thus able to earn a reasonable hourly wage. Our results show that their behavior was not driven by the desire to minimize unpaid time but that they expended real effort on the experimental tasks.

[20] One participant closed the browser after two rounds, could not further participate in the experiment and was therefore excluded from the analysis. After the participant quit the experiment, their group continued with the remaining four group members. Thus, our analysis for BT is based on nine groups with 5 members per group and one group with four members. Participants in the reduced group only saw the fifth participant is always writing empty reviews and therefore had the same information as the participants in the group of five.

[21] 86 participants rated the maximum of 20 reviews. 10 participants quit early; one of them rated one review, seven of them rated five reviews each, one of them rated 10 reviews, and one of them rated 15 reviews. We asked them to complete all reviewers but some quit earlier due to survey being run online.

[22] We use the review length as a proxy for quality for several reasons. Mudambi and Schuff (2010) have shown that review length has a positive effect on quality, because longer reviews include more detailed

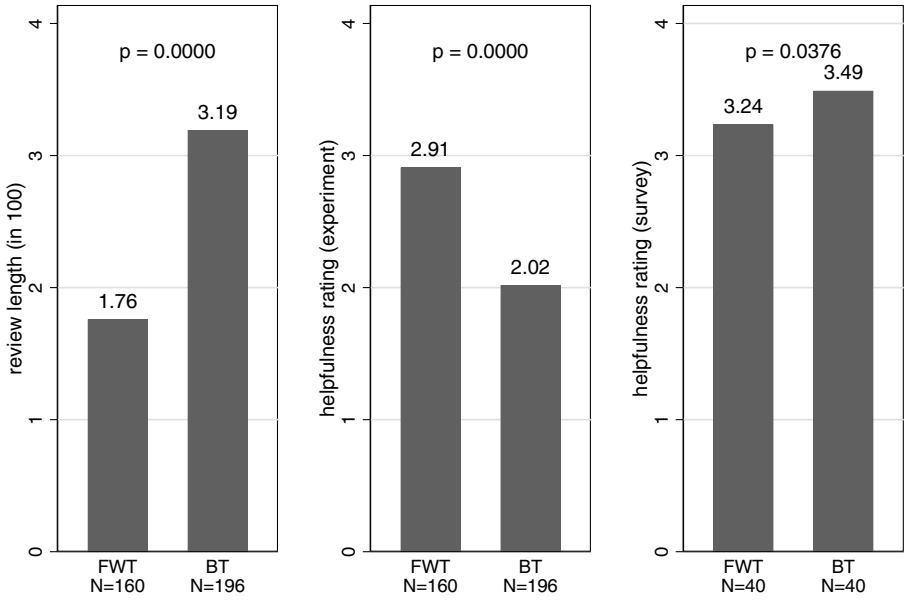

Fig. 4 Mean values for review length, helpfulness rating from the experiment, and helpfulness rating from the survey (*P*-values in graphs are from two-sided Mann-Whitney tests. Data are shown for all four rounds)

Fig. 4 summarizes the pooled data. We have four observations per participant (one observation for each round), resulting in $4 \times 40 = 160$ observations in FWT and $4 \times 49 = 196$ observations in BT. In BT, reviews are longer, but receive lower helpfulness ratings. The presumably unbiased helpfulness ratings from the survey do not indicate that BT reviews are less helpful. This is an indicator of strategic downvoting in the experiment.

First, we compare helpfulness ratings in the experiment. In BT, the mean is lower (2.02 in comparison to 2.91 in FWT, two-sided Mann-Whitney test, $z = 8.001$, $p = 0.0000$).[23]

The comparison of these means, however, does not consider the quality of the reviews. Lower helpfulness ratings in BT might be justified if the review quality is lower. This is unlikely because reviews are significantly longer (216 characters in FWT and 302 in BT, two-sided Mann-Whitney test, $z = -3.202$, $p = 0.0014$). More importantly, we can confront this with the data from the survey. For each treatment, we have 40 randomly selected reviews which were evaluated by external survey

descriptions of the product. Korfiatis et al. (2012) also find that helpfulness of reviews is increased with length. As participants in our experiment had to pay for each character longer reviews come with higher monetary opportunity costs of 0.4 cents per character. In addition, review length and writing time are correlated. We performed the result analysis with review writing time as well and get qualitatively similar results.

[23] To check robustness, we repeat this test using data from the first round only because then, observations are independent (40 obs. for FWT, 49 obs. for BT). In BT, the mean is lower (2.56 in comparison to 3.04 in FWT, two-sided Mann-Whitney test, $z = 2.972$, $p = 0.0030$).

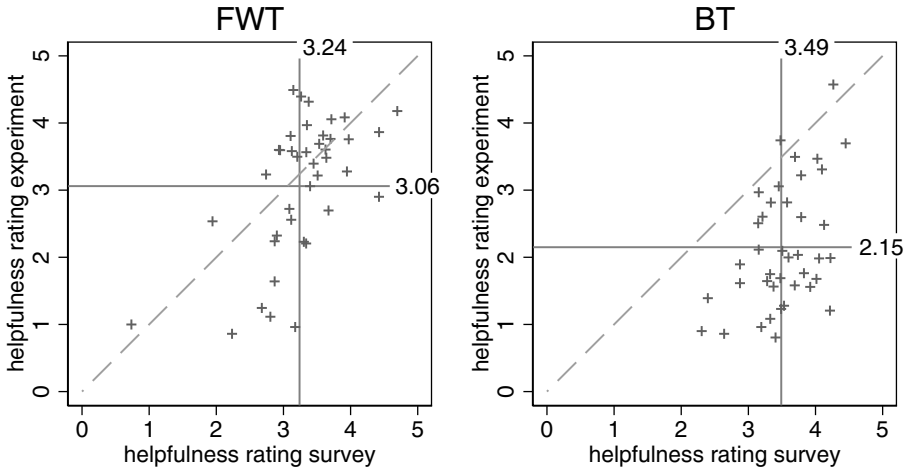

Fig. 5 Average helpfulness ratings in the experiment and in the survey by treatment (Each review is represented as a point. The point's y-coordinate is given by the mean of all average helpfulness ratings received by the specific review in the experiment. A point's x-coordinate is given by the average of all helpfulness ratings received by the specific review in the survey. Horizontal and vertical lines represent mean values based on the average helpfulness ratings assigned to all reviews that were rated both by participants of the experiment and by survey participants)

participants (see Sect. 3.2). Judged by the unbiased reviews from the survey, average review quality is higher in BT (3.17 in FWT and 3.56 in BT, $z = -2.156$, $p = 0.0311$, two-sided Mann-Whitney test). Hence, we can exclude the possibility that lower helpfulness ratings in BT are driven by the quality of the reviews.

4.1.1 Result 1

There is strategic downvoting in treatment BT. We find clear evidence for Hypothesis H1.

Fig. 5 illustrates Result 1 graphically by plotting the mean helpfulness ratings each review received in the experiment against the rating it received in the survey. Points above (below) the 45-degree line indicate reviews for which the helpfulness rating in the survey was lower (higher) than the helpfulness rating in the experiment. In FWT, all points are distributed around the 45-degree line: 18 points lie above, 19 below, and 3 exactly on the 45-degree line. This is very different in BT, where only 3 points lie above, and 37 points lie below the 45-degree line.

To further validate result 1, we run a linear regression[24] of the helpfulness rating on review length, an indicator variable for BT, and the interaction between both independent variables (see first column in Table 1). The review length positively

[24] Results are robust with respect to the model specification. Order-logit regressions or random effects regressions yield qualitatively identical results. For sake of simplicity we only report the OLS estimates. For robustness reasons, we performed a FE-regression which yields similar results and can be found in the appendix.

Table 1 OLS regressions with average helpfulness rating as dependent variable

	(1)	(2)
Review length	0.629***	0.583***
	(0.074)	(0.067)
Bonus treatment	–0.568***	–0.701***
	(0.182)	(0.169)
Bonus treatment * review length	–0.384***	–0.321***
	(0.096)	(0.090)
Round = 2	–	–0.132
		(0.117)
Round = 3	–	–0.388**
		(0.162)
Round = 4	–	–0.571***
		(0.189)
Constant	1.803***	2.157***
	(0.058)	(0.150)
Observations	356	356
R^2	0.486	0.534
Adjusted R^2	0.482	0.526

Data from 89 participants for four rounds
Clustered standard errors by participant and group in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

impacts the helpfulness rating. The coefficient for the indicator variable for treatment BT is negative and significant. Also, the coefficient for the interaction term (BT * review length) is negative and significant. Hence, an increase in review quality leads to an increase in the helpfulness rating, but the increase is smaller in BT. In the second regression in Table 1, we add controls for rounds (taking round 1 as baseline). Note that the indicator variables for rounds do not only capture the effect of learning, but also individual product characteristics, as participants rated a different product each round. Our results are robust to the inclusion of round effects.

The evidence described in the previous paragraphs indicates that helpfulness ratings differ between treatments. Helpfulness ratings expressed by participants in BT are clearly biased downwards. Hence, they do not reflect the true underlying quality but are driven by strategic downvoting.[25]

As we cannot reject the H1 (strategic downvoting), we analyze the evidence related to the quality of the reviews. The comparison of helpfulness ratings from the survey reveals that review quality is higher in BT (see above). Contrary to our expectations, the data does not provide support for H2.

4.1.2 Result 2

Despite strategic downvoting, the average quality of reviews is higher in treatment BT.

[25] In Appendix B we provide a further regression on the assignment of helpfulness votes depending on review length, strategic considerations and demographics.

Table 2 Since review length is truncated in the interval [0, 400], we performed Tobit regressions with review length as dependent variable

	Flat wage treatment	Bonus treatment
Helpfulness rating (t–1)	0.491*** (0.0904)	0.335*** (0.117)
Review length others (t–1)	0.164*** (0.0310)	0.123*** (0.0310)
Won reward (t–1)	–	–0.153 (0.212)
WTP	0.0153 (0.0732)	0.0713 (0.0775)
Score	0.151** (0.0744)	0.0662 (0.0525)
Econ	–0.452* (0.263)	0.149 (0.199)
Gender	0.402* (0.226)	0.0401 (0.144)
Round = 3	–0.397* (0.224)	–0.267 (0.194)
Round = 4	–0.205 (0.223)	0.0282 (0.199)
Constant	–1.627*** (0.496)	0.869** (0.415)
Pseudo-R^2	0.17 (0.119)	0.09 (0.0741)
Observations	120	147

In FWT, data from 40 participants, rounds 2–4
In BT, data from 49 participants, rounds 2–4
Robust standard errors (by subject) in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

If the quality of reviews is not affected by strategic downvoting, the question remains which factors do influence the quality of reviews. To shed some light on this question, we performed Tobit regressions with review length as the dependent variable (see Table 2).

Higher helpfulness ratings in a given round increase review quality in the following round. This positive correlation indicates that helpfulness ratings may act as approval, similar to Masclet et al. (2003), Dugar (2013) and Greiff and Paetzel (2015).

A similar effect arises due to the dynamics within a group. If all other group members provide high-quality reviews, participants also increase the quality of their reviews. These self-reinforcing effects are similar to the coordinating effect of high contributions often observed in public good games (e.g., Weimann 1994).

We also examined the influence of winning the bonus on review behavior in the round following the win. Winning the bonus had a negative but non-significant effect on review quality.

We included controls for the product score and the willingness to pay (WTP). Only in FWT, the score has a weakly significant effect, indicating that in this treatment, participants who perceive the product as better tend to write longer reviews.

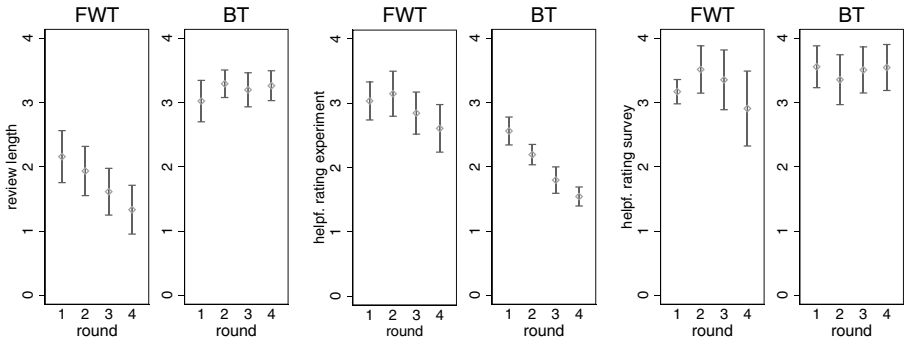

Fig. 6 Mean values and 95% confidence intervals by round

Controlling for gender and economics-related major showed that reviews written by female participants were longer in FWT but not in BT. Participants with an economics-related major wrote shorter reviews in FWT. But effects are only weakly significant.

4.2 Behavior Over Time

Fig. 6 depicts our three main variables by round. These figures provide support for an additional Result 3.

4.2.1 Result 3

Over time, review quality decreases in FWT but not in BT. Strategic downvoting becomes more severe over time.

Similarly to repeated public goods games, in FWT, we observe that the number of characters written decreases over time. In BT, the length of reviews does not decrease but stays constant. It seems that the existence of the bonus prevents a decrease in review quality. Participants do not shy away from writing long and qualitative reviews in spite of their reviews being voted down strategically. Learning does not play a role here as we find these long reviews until round 4 in BT. In Table 2, we report dummies for later rounds. Only for FWT they are overall negative and one of them gets significant while for the BT they are not significant at all. This support the results found in Fig. 5[26].

The pattern is mirrored if we look at helpfulness ratings from the survey. In FWT, these ratings show a downward trend while in BT they stay constant. To the contrary, the helpfulness ratings in the experiment show a very different pattern. While in FWT they correlate with the length of the review, in BT the helpfulness ratings sharply decrease over time instead of the length of reviews being high and constant, which again confirms our hypothesis on strategic downvoting.

[26] While in BT the review length does not decrease with the number of the round, in FWT it decreases with 30 characters per round. This is shown in Table B.3 in appendix B.

5 Discussion and Conclusion

Both quantity and quality of customer-written product reviews have positive impacts on purchase intentions, sales, customer satisfaction, and welfare. Retailers and online platforms use different incentives to increase review quantity and quality. In our experiment, we focus on review quality and compare two different incentive schemes. Under the first incentive scheme, reviewers receive a flat salary per review, independent of the review's quality. Under the second incentive scheme, only the reviewer who wrote the highest-quality review receives a bonus.

Under both incentive schemes, helpfulness ratings are assigned by the other reviewers. Theory predicts that the bonus will lead to strategic downvoting. Reviewers will assign low helpfulness ratings to reviews written by others, because this maximizes the chances of winning the bonus. If reviewers anticipate strategic downvoting, the quality of reviews might deteriorate because of crowding-out effects and because helpfulness ratings do not express approval.

Our data shows that the quality-contingent bonus indeed leads to strategic downvoting. Although the data provides clear evidence for strategic downvoting, the bonus does not have a negative effect on review quality. Review quality remains constant in the presence of the bonus scheme but decreases over time when reviewers receive a flat salary.

We chose our two incentive schemes such that the expected monetary payoffs are identical. This allows us to rule out differences in expected payoffs as an explanation for the observed differences in review quality. Given that most retailers do not reward reviews with a fixed payment and most reviewers are paid nothing, we expect the difference in review quality to be even larger in the real world.[27]

In our experiment, all participants write reviews and rate the quality of others' reviews. In reality, there are four mutually exclusive roles: customers who write reviews and vote on the quality of reviews ("reviewers" as in our experiment), customers who write reviews and do not vote on the quality of reviews, customers who do not write reviews but vote on the quality of reviews ("voters"), and silent customers who neither write reviews nor judge the reviews' quality ("silent customers"). Only "reviewers" have an incentive for strategic downvoting, but they are the minority, which may raise the concern that this could change our predictions. If the majority of votes are cast non-strategically, the impact of strategic downvoting may be quite small and strategic downvoting might not be a problem. However, this is not the case. Consider two reviews written by "reviewers", who compete for a bonus given to the most helpful review. Assuming that the votes cast by "voters" are unbiased, both reviews will receive the same number of positive and negative helpfulness ratings from "voters". The helpfulness ratings assigned by "reviewers" themselves will be decisive for determining who gets the bonus. The impact of strategic downvoting might be a problem in reality, even though "reviewers" are the minority.

[27] More precisely, one could argue that the fixed payment results in a level-effect. Without any payment, review quality would be lower than with the fixed payment, but in both cases, review quality would decrease over time.

Our results have direct managerial implications for retailers. First, tournament incentivize schemes have no adverse effects on review quality. This suggests that these incentive schemes increase the amount of pre-purchase information. Because only a small number of products receive professional reviews (e.g., reviews in big newspapers) but a much larger number of products are reviewed by customers, employing tournament incentive schemes to generate pre-purchase information is advisable, especially for retailers selling niche products. In addition, our results show that reviewers maintain the high quality even over several rounds. Note, however, that in reality, a tournament incentive schemes could reduce the quantity of reviews written (which was fixed in our experiment), so that there is a tradeoff between the positive effect on quality and the negative effect on quantity.

Second, when determining which reviewers will receive a bonus, these retailers have to respect the fact that helpfulness ratings could be biased. It is therefore desirable that this bias is mitigated. Instead of using the arithmetic average for aggregating helpfulness ratings, retailers could switch to alternative approaches which assign less weight to helpfulness ratings from reviewers which are likely to be motivated to downvote other reviewers. Possibly, one could employ statistical methods or machine learning to estimate the size of the bias, and use then this data to devise an aggregation procedure which captures the maximum of information from helpfulness ratings (see also Dai et al. 2018).

Third, a closely related problem is the loss in signaling power, which arises from strategic downvoting. If customers cannot rely on helpfulness ratings to help them find high-quality reviews, their search and evaluation costs will increase if they realize this bias. If not, they base their decision on inferior reviews and may regret their purchase decision. Retailers can counteract the loss in signaling power because only reviewers have an incentive for strategic downvoting. "Voters", who never write reviews, have no incentive to do so. By focusing on helpfulness ratings assigned by these customers, retailers can identify the reviews that are most helpful (based on unbiased ratings). In contrast to the machine learning approach above, the disadvantage may be to lose votes, which is especially problematic when new products are launched, and only little votes are gathered. Another option is therefore not to exclude but to mark those helpfulness ratings which came from other reviewers as such.

A fourth implication concerns the relation between the problem of obtaining high-quality reviews and the provision of public goods. In both cases, the benefits are publicly available while costs are private. Our study indicates that a monetary bonus given to the participant who made the highest contribution increases efficiency, even though the "best" contributor is determined endogeneously, which could give rise to strategic downvoting. However, caution should be exercised when generalizing the results from our study to other public good style situations. We have analyzed a market for reviews where each and every review receives exactly the same number of helpfulness ratings from all other participants. Moreover, there are no opportunity costs of assigning helpfulness ratings. Because of these differences, our study might not adequately capture many features of "real-world" public good style situations. It would be worthwhile to analyze how the presence of opportunity costs affect the assignment of helpfulness ratings and consequently, contributions. A further

limitation of this study is the short-run examination of review behavior. It remains unclear whether a bonus still has no negative effect on review quality when review behavior is observed over an extended period of time. These aspects are beyond the scope of our paper, which had the more modest goal of identifying whether a monetary bonus affects the assignment of helpfulness ratings and participants' review writing behavior.

The results derived from this study open up new and interesting questions for future research on how different incentive schemes affect review quality. Future research could use more complex experimental designs to analyze the changes discussed in the previous paragraph. In addition, field data could be used to shed some light on the size of the "downvoting" bias in existing review systems and could develop and test alternative helpfulness-based incentives that do not give rise to strategic downvoting.

Acknowledgements We would like to thank Janis Cloos, Johann Graf Lambsdorff, Susanna Grundmann, Philipp Krügel, Thomas Niemand, Fabian Paetzel, Anna Ressi, Manuel Schubert, Katharina Werner and two anonymous reviewers for their helpful comments.

Funding The research was financed by the Gesellschaft für experimentelle Wirtschaftsforschung e. V. (GfeW) with the Heinz-Sauermann sponsorship award.
Funding Open Access funding provided by Projekt DEAL.

Appendix A

Instructions

Oral Instructions (Read Aloud Before the Experiment, in German)

Welcome to the experiment and thank you very much for your participation. I will briefly read you some general explanations about the experiment. Please do not click on "Start experiment" until the end of these instructions. The participants of the experiment are all here in this room and are all taking part in the same experiment. The experiment aims at gaining insights on human behavior. The experiment lasts about 90 min and on average you will receive between 7 and 15 €, depending on your behavior, but at least 2 €. You will play anonymously and can't coordinate with each other. The disbursement of payoffs will also be carried out anonymously. No other participant will see how much you receive and the experimenters will not find this out either. During the experiment you may have to wait for the other participants. This may take a few minutes. Please remain patient during this time.

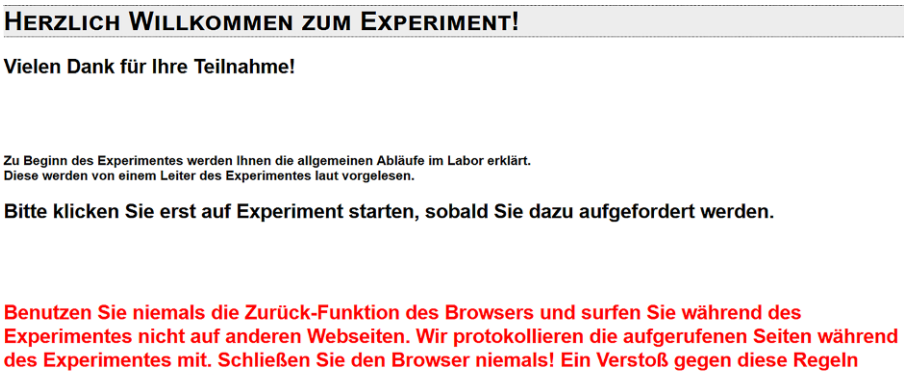

Fig. A.1 Instructions—Welcome to the experiment!

When everyone has finished the experiment, you will be asked to go outside one after the other. There you will receive your payment. All instructions and explanations can be found on the following screen pages. Please read all the information carefully before leaving a screen by mouse click. Once you leave a screen, you will not be able to access it again. Never use the back function of the browser and do not surf on other websites during the experiment. We will log the accessed pages during the experiment. Never close the browser! A violation of these rules will exclude you from any payoff. Please remain calmly seated at your workplace. Please refrain from any conversations. If you have any questions, please raise your hand. We will then come to you. Now click on "Start experiment".

Instructions—Welcome to the Experiment!

Thank you very much for your participation! At the beginning of the experiment the general laboratory procedures will be explained. These will be read aloud by the experimenter. Please click on "Start experiment" as soon as you are asked to do so. Never use the back-function of the browser and do not surf on other websites during the experiment. We log the accessed pages during the experiment. Never close the browser! A violation of these rules will exclude you from any payoff (Fig. A.1).

Instructions—General Information

In this experiment you have the task to write product reviews in an online shop and to evaluate the reviews of 4 other reviewers. You always interact with the same 4 reviewers over 4 rounds throughout the experiment. Each reviewer has identical tasks and receives the same instructions. At the beginning of the experiment you will receive € 6.4 as initial endowment. In each round you will receive earnings or deductions to your payoff account depending on your behavior. The remaining payoff account at the end of the experiment will be paid to you. In the experiment characters will be converted into Euro. 25 characters = 10 Eurocents (Fig. A.2).

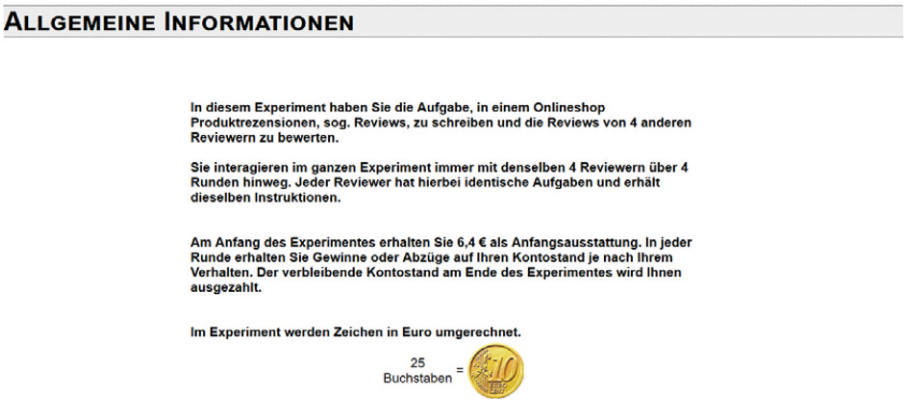

Fig. A.2 Instructions—General information

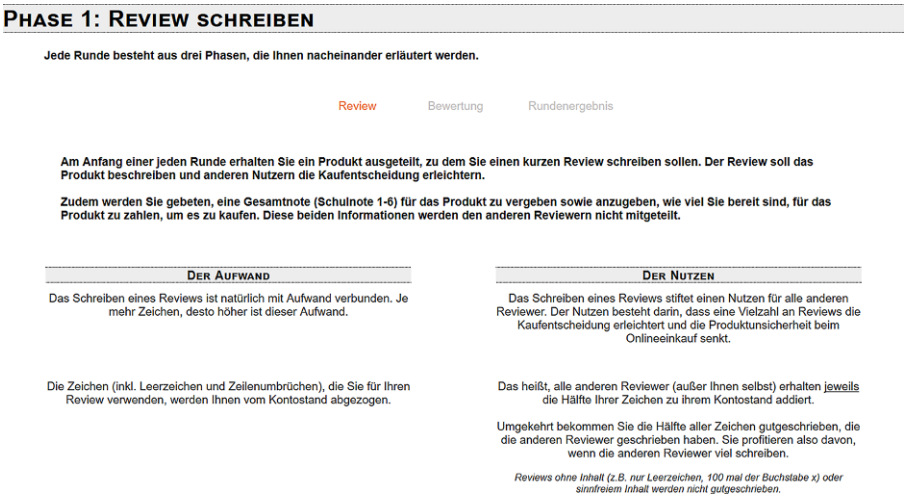

Fig. A.3 Instructions—Phase 1: Writing the review

Instructions—Phase 1: Writing the Review

Each round consists of three phases, which are explained to you one after the other.

Review—Rating—Round result

At the beginning of each round you will receive a product to which you should write a short review. The review should describe the product and make it easier for other users to make a purchase decision.

You will also be asked to give the product an overall grade (school grade 1–6) and indicate how much you are willing to pay to buy the product. This information will not be shared with other reviewers (Fig. A.3 and Table A.1).

Table A.1 Instructions—Phase 1: Writing the review

The effort	The benefit
Of course, writing a review involves a lot of effort. The more characters, the higher the effort. The characters (including spaces and line breaks) you use for your review will be deducted from your payoff account.	Writing a review provides a benefit for all other reviewers. The benefit lies in the fact that a large number of reviews facilitates the purchase decision and reduces product uncertainty when buying online. This means that all other reviewers (except you) each receive half of their characters added to their account balance.
Example:	In the same way, half of all characters written by the other reviewers will be credited to your account. So you benefit if the other reviewers write a lot.
You write 0 characters → deduction: € 0	Reviews without content (e.g. only spaces, 100 times the letter x) or meaningless content will not be credited.
You write 400 characters (maximum) → deduction: € 1.6	Example:
	You write 400 characters (your deduction: € 1.6)
	→ credit 200 characters (= € 0.8) for each other reviewer
	→ credit characters for all others together: € 3.2

Instructions—Phase 2: Rating Reviews of Others

After writing your own review you will be asked to evaluate the reviews of the 4 other participants regarding their helpfulness.

Helpfulness Assessment Helpfulness is the usefulness of the review for a possible purchase decision. You can rate helpfulness on a scale of 5 from unhelpful (1 star) to very helpful (5 stars).

Please note that all reviews must be rated (at least 1 star = not helpful at all).

ONLY IN TREATMENT BT Award for the review with the best helpfulness rating.

The review that gets the best average helpfulness rating from the other 4 reviewers gets a price of € 5 in addition to its normal payoff.

If several reviews have the same average rating, the price will be split.

ONLY IN TREATMENT FWT The rating has no effect on the payoffs (Fig. A.4).

Instructions—Phase 3: Result of the Round

At the end of each round, you will be informed about your total payoff for that round and the average helpfulness rating for your own review.

ONLY IN TREATMENT 1 You will also receive € 1 for evaluating the product (overall impression, willingness to pay) and rating the other reviews.

Your Payoff at the End of the Round You have written a review with X characters. These characters will be deducted from your account balance. The other 4 reviewers wrote reviews with a total of Y characters.

Since you also benefit from the reviews as a user, you will receive half of the characters credited as a payout.

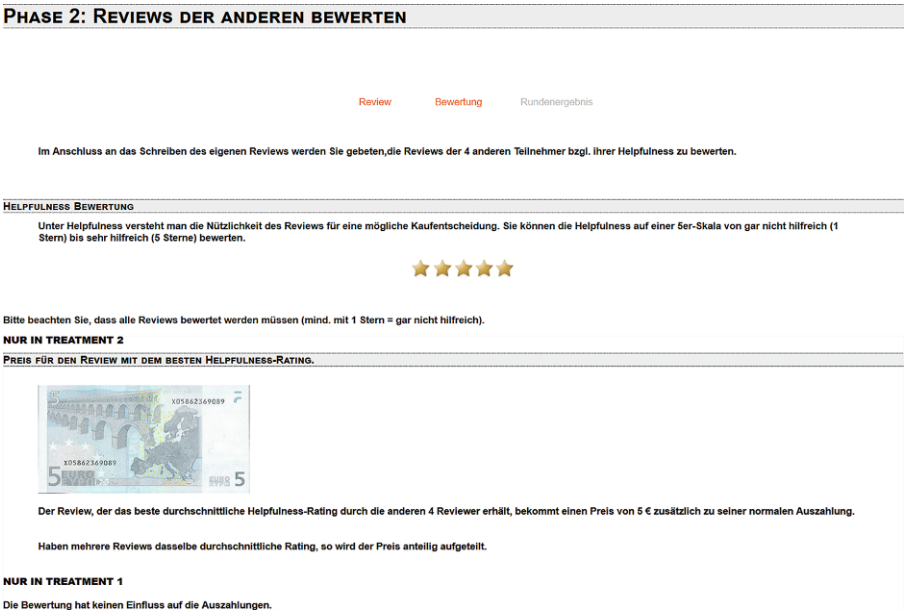

Fig. A.4 Instructions—Phase 2: Rating reviews of others

ONLY IN TREATMENT FWT You will additionally receive 1 € credited for evaluating the product and rating the other reviews.

ONLY FOR THE BEST REVIEW IN TREATMENT BT You have the best helpfulness rating and get an additional payoff of € 5.

Purchase of Test Products In the first phase you indicated how much you are willing to pay for the product. If this willingness to pay is higher than the purchase price of the product (purchase on the Internet), you will receive the product at purchase price and the purchase price will be deducted from your payoff account. Please note that you will receive the product at purchase price. Important: The indication of the willingness to pay is binding.

Example If you are willing to pay a high price, you can enter it as willingness to pay, but you will then receive the product at the cheaper purchase price.

Payment and Questionnaire At the end you will find an anonymous questionnaire about your experiences with online review systems.

When you leave the laboratory, you will receive your payment and, if applicable, the products you have purchased (in their original packaging) (Fig. A.5).

Decision Screens

All decision screens show a bar with (1) the round number, (2) the stage in the current round (review, rating, result of the round), (3) the total payoff (excluding

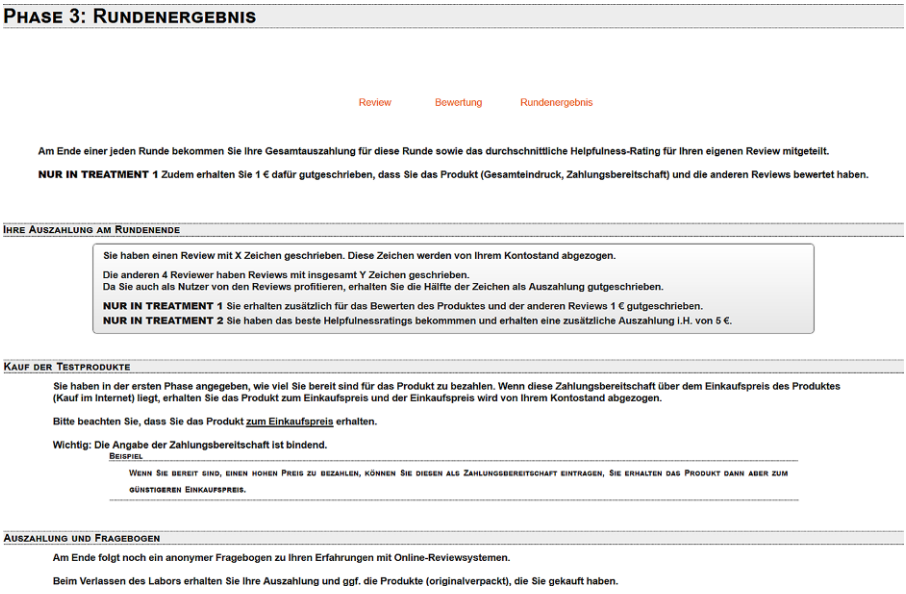

Fig. A.5 Instructions—Phase 3: Result of the round

Fig. A.6 Decision screens

the current round) and (4) the conversion rate of characters into Euros (25 characters = 10 Eurocents) (Fig. A.6).

Decision Screen 1—Writing Review

On the first screen, participants can write their review. The actual text on screen is marked by "". The remaining text is explanation (Fig. A.7).

1. General instructions for writing the review: "The product is now distributed. It is collected after the round. Please write a review about the product. You have max 400 character available. Each character that you write will be deducted from your payoff account. By writing the review, you create utility for the other reviewers. They each get half of your characters as payoff."
2. In the text field, subjects could write their review and were informed about the remaining characters.
3. In the info field, subjects could see (in live) how writing affected their payoff. The text states: "Your payoff account (after deducting the characters from your review)".

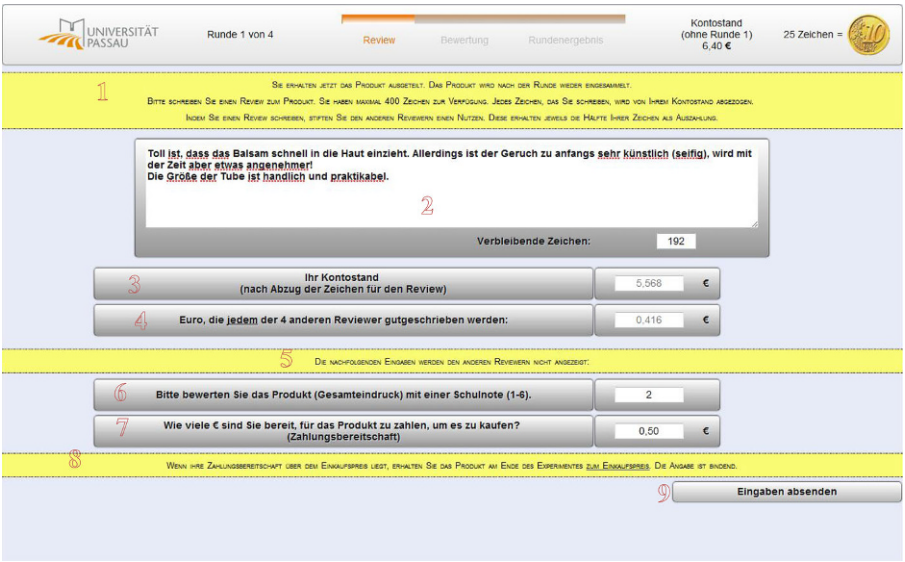

Fig. A.7 Decision Screen 1—Writing review

4. In the info field, subjects could see (in live) how writing affected the others' payoff. The text states: "Euros that will be given to *each* of the 4 other reviewers.".
5. This message informs subjects about the fact that "The following inputs are not shown to the other reviewers". (meaning 6 and 7).
6. Subject have to rate the product. "Please rate your overall impression of the product with a school grade (1–6)". In Germany, 1 is the best and 6 the worst grade. You pass with at least a 4.
7. Subjects have to express their willingness to pay for the product. "How many € are you ready to pay to buy the product? (willingness to pay)".
8. The message at the bottom informs subjects about the binding nature of the input in 7. "If your willingness to pay is above the buying price of the product, you get the product at the end of the experiment *for the buying price*. This input is binding."
9. "Submit inputs"

Decision Screen 2—Rating Reviews

Then subjects rate the reviews of the 4 other reviewers (Fig. A.8).

1. The instruction state "Please rate the helpfulness of the 4 other reviews with a star rating. (1 = not helpful at all, 2 = little helpful, 3 = average helpful, 4 = helpful, 5 = very helpful). The reviewer with the highest average rating in your group of 5 gets additional 5 €. With a tie, the award will be split". (for the BT treatment). In the FT, the last two sentences were changed to "The rating does not influence your payoff." The € 5 bill was only shown in the BT treatment.

Fig. A.8 Decision screen 2—Rating reviews

2. Each review is headed by stating "Participant A has written the following review:"
3. The review was reproduced. The number of characters was added in brackets "(This review has 347 characters)".
4. "Please rate the helpfulness of the review".

Decision Screen 3—Feedback

The third screen provides feedback (Fig. A.9).

1. In the BT treatment, subjects were informed if they had the highest rating. "You have the highest helpfulness rating and get € 5." If not they were informed as well. In the FWT treatment, this part was omitted.
2. Subjects were informed about the average helpfulness rating they obtainted. "Your review got an average helpfulness rating of 1.25 by the 4 other reviewers".
3. Feedback on payoffs was provided: "In this round, you received the following payoff. You wrote a review with 208 characters. These characters were deducted from your payoff account. The other reviewers wrote reviews with 359 characters in total. As a user, you profit from these reviews and you get half of all characters as payoff. You had the highest helpfulness rating and got an additional payoff of € 5." The last sentence was changed accordingly in treatment FT to "You will additionally receive 1 € for evaluating the product and rating the other reviews."

Fig. A.9 Decision Screen 3—Feedback

4. Feedback on total payoff. "Overall, in this round your payoff has increased by € 4.886".
5. Button to start the next round with warning "Please click only on the button once you read all instructions".

Questionnaire After Experiment

Thank you for participating in the experiment! The following questionnaire will collect your experiences with online product reviews in the "real world". Please fill in the following questionnaire carefully. Answering the questions takes about 10–15 min. The data collected cannot be personally attributed to you. This questionnaire has 29 questions.

Experience as Reviewer—Part I

The following questions ask about your experience when writing online product reviews.

1. How many product reviews have you written on Amazon?
   - None
   - Up to 10
   - Up to 20
   - Up to 50
   - More

2. Have you already written reviews for other e-shops than Amazon? *Question only asked if product review written (question 1).*
   - Yes
   - No
3. When was the last time you wrote a review? *Question only asked if product review written (question 1).*
   - In the last week
   - In the last month
   - In the last year
   - Over a year ago

Motivation Reviews in General

4. What motivates you to write product reviews? *Question only asked if product review written (question 1).* Please select the appropriate answer for each item: (1 = absolutely true, 2 = true, 3 = hardly true, 4 = not true)
   - In return, I'm receiving monetary incentives such as money or coupons.
   - I will receive free test products in return.
   - My experience helps other customers with the assessment of product quality.
   - I enjoy it to write reviews.
   - I like to be in contact with other reviewers and readers.
   - A reputation as a good reviewer is important for me.
   - I want to support other customers to buy the right products.
   - I like to exchange myself with people who have similar interests.
   - I hope that others reviewers and readers will give me advice on problems with products.

Experience as Reviewer—Part II

5. What was (or is) your best reviewer rank on Amazon? *Question only asked if product review written (question 1).*
   - Best 100
   - Best 1000
   - Best 10,000
   - Best 50,000
   - Best 100,000
   - Higher than 100,000
   - I do not know
6. Do you participate in the Amazon Vine program? *Question only asked if product review written (question 1).*
   - Yes
   - No

7. Have you already received "Not helpful" votes for your reviews? *Question only asked if product review written (question 1).*
   - Never
   - Few
   - Some
   - A lot
   - I do not know
8. Were you able to understand this evaluation? *Question only asked if product review written (question 1) and if not helpful votes received (question 7).*
   - All
   - Almost all
   - Some
   - Few
   - None
9. What are the reasons for you to write a positive product review? *Question only asked if product review written (question 1).* Please select the appropriate answer for each item: (1 = absolutely true, 2 = true, 3 = hardly true, 4 = not true)
   - I want to help other people with my positive experience.
   - I want to give other people the opportunity to buy the right product.
   - So I can express the joy of a good purchase.
   - I like to tell other people about a successful purchase.
   - So I can show other people that I have bought cleverly.
   - I would like to recommend the company.
   - I like to support good companies.
10. What are the reasons for you to write a negative product review? *Question only asked if product review written (question 1).* Please select the appropriate answer for each item: (1 = absolutely true, 2 = true, 3 = hardly true, 4 = not true)
    - This is how I better process the frustration over a bad buy.
    - So my anger over a bad buy is reduced faster.
    - I want to warn other customers about bad products.
    - I want to spare other customers a bad product experience.
    - I want to pay back the manufacturer of the bad product.
    - It is less time-consuming than complaining to the manufacturer by phone or e-mail.
    - I believe that the manufacturer will solve the problems of their product faster if I discuss them publicly.
    - The provider of the review platform will forward my complaint to the right place at the manufacturer.
11. How often do you shop on Amazon?
    - More than 30 products a year
    - Between 30 and 10 products a year
    - Less than 10 products a year
    - Never

12. How often do you shop on the Internet?
    - More than 30 products a year
    - Between 30 and 10 products a year
    - Less than 10 products a year
    - Never
13. How often do you read product reviews on Amazon before you make a purchase of a product?
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never
14. How often do you read product reviews on other online platforms before you make a purchase of a product? *Question only asked if customer never read on Amazon (question 13).*
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never
15. I read first those reviews which ... *Question only asked if review are read (question 13).* Please select the appropriate answer for each item: (1 = always, 2 = often, 3 = sometimes, 4 = rarely, 5 = never)
    - is displayed first automatically
    - was rated as most helpful by other customers
    - is the most recent
    - evaluate the product best
    - evaluate the product worst
16. Before I decide to purchase a product, I read all the available product reviews (on Amazon).
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never

Reading Motivation Customers

17. I read the product reviews of other customers ... *Question only asked if reviews read (question 13).* Please select the appropriate answer for each item: (1 = absolutely true, 2 = true, 3 = hardly true, 4 = not true)
    - because they're helping to make the right purchase decision.
    - because it saves a lot of time if I want to inform me about a product before buying it.

- in order to find advice and solutions for problems.
- because I feel better, when I read that other people have the same problem with a product.
- to benefit from the experiences of others, before I buy a product.
- because it is the fastest way to get information about a product.
- because I get to know about recent trends.
- to find confirmation, that I have bought the right product.
- to compare my product evaluation with that of other people's.
- because I like to share the experiences with others reviewer.
- because I am rewared for reading and rating (e.g. vouchers, free test products).
- to find the right answers when I have problems with the product.
- because I like to be part of the review community.
- because I am interested in new products.
- to find out if I am the only one with a certain opinion about a product.

18. How often have you rated the reviews of other customers? *Question only asked if reviews read (question 13).*
    - Never
    - Up to 10 times
    - Up to 20 times
    - Up to 100 times
    - More often

Customer View Helpfulness

19. I find a customer review particularly helpful if it ... *Question only asked if reviews read (question 13).* Please select the appropriate answer for each item: (1 = absolutely true, 2 = true, 3 = hardly true, 4 = not true)
    - Discusses the disadvantages of the product.
    - Discusses the advantages of the product.
    - Is easy to read.
    - Contains much expert information.
    - Discusses both advantages and disadvantages of the product.
    - Is very short.
    - Extensively discusses the experiences of the reviewer with the product.

Customer Opinion Formation

20. I find reviews that are already rated as "helpful" by many other customers. *Question only asked if reviews read (question 13).*
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never

21. Have you ever given up buying a product because there have been no or very few customer reviews?
    - Yes
    - No
22. Have you ever bought a product, although you had found mostly bad reviews for it before the purchase?
    - Yes
    - No
23. If a product in which I am interested is rated negatively in a review, I do not make a purchase.
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never
24. If a product in which I am interested is rated positively in a review, I buy the product or at least seriously consider a purchase.
    - Always
    - Often
    - Sometimes
    - Rarely
    - Never

Demographic Information

25. Are you male or female?
26. Are you enrolled as a student?
    - Yes
    - No
27. Which subject do you study? *Question asked only if enrolled as student (question 26).*
    - Teaching
    - Law
    - IT
    - Economics
    - Business Administration
    - Media and Communication
    - International Cultural and Business Studies
    - European Studies
    - Governance
    - Internet Computing
    - Mobile and Embedded Systems
    - Linguistics and Text Sciences
    - Other

28. In which subject area did you achieve your highest level of education? *Question asked only if not enrolled as student (question 26). Options are the same as in question 26.*
29. How old are you?

Appendix B

Additional Results

What are the factors that drive the assignment of helpfulness ratings? To answer this question, we performed an OLS regression for each treatment, with the helpfulness rating as the dependent variable (see Table B.1).[28] The independent variables fall into four categories.

The first category refers to *review quality* and includes only review length, our proxy for quality. In FWT, review length is positively and highly correlated with helpfulness. Helpfulness increases by 0.843 units for 100 additional characters; in BT, the increase is only 0.157 (and significantly lower than 0.843, two-sided *t*-test, $p < 0.001$).

The second category contains all variables which reflect *strategic considerations*. Due to strategic considerations, the length of a participant's own review could be positively correlated with the helpfulness rating. Consider a participant who wrote a short review while her group members wrote long reviews. In order to win the reward, she has to obtain the highest helpfulness rating in her group. If she expects that longer reviews receive higher helpfulness ratings, she expects the other group members' helpfulness ratings to be higher than the helpfulness rating she receives. In such a situation, the only thing she can do to maximize her chances of winning the reward is to assign lower ratings to others. In fact, we expect an inverse relationship between the length of the review written by a participant and the incentive to downvote others. The shorter the own review, the greater incentive to downvote other participants. Contrary to these considerations, we found the length of a participant's own review to have no effect on the helpfulness rating she assigns.

The helpfulness rating received in the previous round could have a positive effect. Such a pattern would be consistent with indirect reciprocity, where participants who have received a high rating in one round (a form of approval) are more likely to assign a high rating in the next round. Our data does not support this pattern. In both treatments, the previous round's helpfulness rating has no effect.

The last strategic factor is the event of winning the bonus. In BT, winning the bonus has no significant effect on helpfulness ratings.

The third category of variables pertains to *product evaluation*. Differences between participants in their perception of a review's helpfulness could be driven by differences in product evaluation. Participants may consider reviews which voice

[28] Performing an ordered probit regression yields qualitatively similar results but with OLS the coefficients are easier to interpret. Results are also robust to a random effects specification.

Table B.1 OLS regressions with helpfulness ratings assigned to other as dependent variable

	Flat wage treatment	Bonus treatment
REVIEW QUALITY		
Review length other	0.843***	0.157**
	(0.0674)	(0.0683)
STRATEGIC CONSIDERATIONS		
Review length own	–0.0176	0.129
	(0.0575)	(0.0981)
Lagged mean rating	0.0280	–0.147
	(0.0675)	(0.163)
Won in last round	–	0.0697
		(0.165)
PRODUCT EVALUATION		
Diff. in scores	–0.0640**	–0.0221
	(0.0273)	(0.0247)
Squared diff. in scores	–0.00892	–0.00930
	(0.0103)	(0.00991)
Diff. in WTP	0.0109	0.0222
	(0.0206)	(0.0454)
Squared diff. in WTP	–0.00321	0.00213
	(0.00253)	(0.0132)
CONTROL VAR		
Econ	0.316	0.0213
	(0.237)	(0.206)
Gender	0.184	0.290
	(0.166)	(0.205)
Constant	1.208***	1.203***
	(0.330)	(0.380)
Controls for groups	YES	YES
Controls for rounds	YES	YES
Observations	480	576
F	18.45	13.16
R^2	0.447	0.265
Adjusted R^2	0.425	0.237

"Review length other" is the number of characters (in 100 s) of the review written by the participant to whom the helpfulness rating is assigned

"Review length own" is the number of characters (in 100 s) of the review written by the participant who assigs the helpfulness rating

Standard errors clusters by participants in parentheses

In FWT, data from 40 participants, rounds 2–4, 4 reviews written per participant and round. In BT, data from 49 participants, rounds 2–4, 4 reviews written per participant and round

For FWT we have 480 observations because we have 40 participants (8 groups with 5 participants per group), 3 rounds (round 2–4), and each participant wrote 4 reviews per round

For BT, we have 49 participants (10 groups with 5 participants per group, minus 2 participants who had to be excluded), 3 rounds (round 2–4), and each participant wrote 4 reviews per round

One participant was excluded because they closed the browser during the experiment and another participant is excluded from this analysis because of missing data for gender

* $p<0.10$, ** $p<0.05$, *** $p<0.01$

different opinions to their own on the product less helpful in a purchase decision. We control for this effect with the variables "product score" and "WTP" (both standardized as *z*-scores). To account for the possibility that only large differences may matter and the direction of the deviation may not, we include squared differences in scores and WTP. Only in FWT, the difference in scores influences the assignment of helpfulness ratings in the expected direction. In BT, different product perceptions does not influence the helpfulness ratings.

The fourth category contains a set of control variables for gender and field of study. None of the control variables had a significant effect. In both regressions in Table B.2 we control for group and round effects.[29]

Summarizing these results, we can say that the assignment of helpfulness ratings is driven by review length. All other factors had only small or non-significant effects. The helpfulness ratings are proxies for review length, which are not biased by strategic considerations or differences in product evaluations. This implies that the differences in correlations between treatments cannot be explained by differences stemming from strategic considerations or differences in product evaluations. Comparing the values for R^2, we see that in FWT compared to BT, the dependent

Table B.2 Fixed effects regressions with average helpfulness rating as dependent variable

	(1)	(2)	(3)
Review length	0.600***	0.483***	0.603***
	(0.0405)	(0.0456)	(0.0347)
Bonus treatment * review length	–0.485***	–0.327***	–0.399***
	(0.0766)	(0.0715)	(0.0641)
Round = 2	–	–0.126	0.247**
		(0.114)	(0.105)
Round = 3	–	–0.402**	0.140
		(0.156)	(0.108)
Round = 4	–	–0.594***	0.0716
		(0.184)	(0.107)
BT * round = 2	–	–	–0.670***
			(0.156)
BT * round = 3	–	–	–0.940***
			(0.207)
BT * round = 4	–	–	–1.139***
			(0.217)
Constant	1.743***	2.043***	1.847***
	(0.119)	(0.138)	(0.101)
Observations	356	356	356
F	111.1	63.25	55.41
R^2	0.260	0.385	0.490
Adjusted R^2	0.256	0.376	0.478

Data from 89 participants for four rounds
Standard errors clustered at group level in parentheses
* $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$

[29] By controlling for group effects, we controlled for the possibility that participants oriented their assignments of helpfulness ratings to reflect the general writing pattern in their group.

Table B.3 Review writing behavior over time with review length as the dependent variable

	Flat wage treatment	Bonus treatment	Flat wage treatment	Bonus treatment
Round	–0.280***	0.0625	–0.305***	0.0532
	(0.0734)	(0.0498)	(0.0669)	(0.0524)
Econ	–	–	–0.683	0.0167
			(0.422)	(0.307)
Gender	–	–	0.296	–0.0726
			(0.279)	(0.217)
Score	–	–	0.148**	–0.00394
			(0.0675)	(0.0459)
WTP	–	–	–0.0939*	–0.0551
			(0.0484)	(0.0623)
Constant	2.459***	3.036***	2.040***	3.168***
	(0.239)	(0.183)	(0.375)	(0.231)
Observations	160	196	160	194

Random effects regressions, standard errors in parentheses
In FWT, data from 40 participants, rounds 1–4. In BT (second column), data from 49 participants, rounds 1–4
In BT (last column), data from 48 participants, rounds 1–4 (one participant excluded because he/she did not indicate his/her gender)
* $p<0.10$, ** $p<0.05$, *** $p<0.01$

variables explain a much larger share of the variance in helpfulness ratings. This is in line with the presence of strategic downvoting.

As a robustness check for our main results in Table 2 we also provide the results from a fixed effect regression, which are well in line with the OLS results.

As a supplement to Table B.1, we regress review length on rounds. In Table B.1, due to the lags, the first round is excluded and we report dummies for round 3 and 4 there. The results from Table B.1 are confirmed by this additional analysis. Only in treatment FWT, the coefficient for the round is negative and significant, indicating that review length decreases by about 30 characters per round.[30] The regressions support the results found in Fig. 5.

References

Bardsley, N., R. Cubitt, G. Loomes, P. Moffatt, C. Starmer, and R. Sugden. 2010. *Experimental economics. Rethinking the rules*. Princeton: Princeton University Press.

Becker, G., M.H. DeGroot, and J. Marschak. 1964. Measuring utility by a single-response sequential method. *Behavioral Sciences* 9(3):226–232.

Benabou, R., and J. Tirole. 2003. Intrinsic and extrinsic motivation. *The Review of Economic Studies* 70(3):489–520.

Berger, J, A.T. Sorensen, and S.J. Rasmussen. 2010. Positive Effects of Negative Publicity: When Negative Reviews Increase Sales. *Marketing Science* 29 (5):815–827

Bickart, B., and R.M. Schindler. 2001. Internet forums as influential sources of consumer information. *Journal of Interactive Marketing* 15:31–40.

Brynjolfsson, E., Y. Hu, and M.D. Smith. 2003. Consumer surplus in the digital economy: estimating the value of increased product variety at online booksellers. *Management Science* 49:1580–1596.

[30] The effect is robust to the inclusion of controls for gender, econ-studies, WTP and score. Here, we report results from random effects regressions with robust standard errors. OLS and Tobit regressions yield similar results.

Chaudhuri, A. 2011. Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14:47–83.
Chen, Y.-H., I.-C. Hsu, and C.-C. Lin. 2010. Website attributes that increase consumer purchase intention: A conjoint analysis. *Journal of Business Research* 63:1007–1014.
Chen, P., S. Dhanasobhon, and M.D. Smith. 2008. All reviews are not created equal: the disaggregate impact of reviews and reviewers at Amazon. http://ssrn.com/abstract=918083. SSRN working paper. Accessed June 8th 2020
Dai, W., G. Jin, J. Lee, and M. Luca. 2018. Aggregation of consumer ratings: an application to Yelp.com. *Quantitative Marketing and Economics* 1(6):289–339.
Davis, D.D., and C.A. Holt. 1993. *Experimental economics*. Princeton: Princeton University Press.
Dechenaux, E., D. Kovenock, and R.M. Sheremeta. 2015. A survey of experimental research on contests, all-pay auctions and tournaments. *Experimental Economics* 18(4):609–669.
Deci, E.L., R. Koestner, and R.M. Ryan. 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin* 125(6):627–668.
Dellarocas, C. 2003. The digitization of word of mouth: promise and challenges of online feedback mechanisms. *Management Science* 49(10):1407–1424.
Dimoka, A., Y. Hong, and P.A. Pavlou. 2012. On product uncertainty in online markets: theory and evidence. *Management Information Systems Quarterly* 36:395–426.
Dugar, S. 2013. Non-monetary incentives and opportunistic behavior: evidence form a laboratory experiment. *Economic Inquiry* 51(2):1374–1388.
Falk, A., and J. Heckman. 2009. Lab experiments are a major source of knowledge in the social sciences. *Science* 326:535–538.
Frey, B.S., and R. Jegen. 2001. Motivation crowding theory. *Journal of Economic Surveys* 15(5):589–611.
Gächter, S., and E. Fehr. 1999. Collective action as a social exchange. *Journal of Economic Behavior & Organization* 39(4):341–369.
Ghose, A., and P.G. Ipeirotis. 2011. Estimating the helpfulness and economic impact of product reviews: mining text and reviewer characteristics. *IEEE Transactions on Knowledge Management and Data Engineering* 23:1498–1512.
Giamattei, M., and Graf J. Lamsbdorff. 2019. classex—an online tool for lab-in-the-field experiments with smartphones. *Journal of Behavioral and Experimental Finance* 22:223–231.
Gill, M., P.F. Evans, V. Sehgal, and M. Da Costa. 2012. *European online retail forecast: 2011 to 2016. Online retail sales growth in Europe will continue to outpace offline growth for years to come*
Gneezy, U., S. Meier, and P. Rey-Biel. 2011. When and why incentives (don't) work to modify behavior. *Journal of Economic Perspectives* 25(4):191–209.
Greiff, M., and F. Paetzel. 2015. Incomplete information strengthens the effectiveness of social approval. *Economic Inquiry* 53(1):5567–5573.
Harbring, C., and B. Irlenbusch. 2011. Sabotage in tournaments: evidence from a laboratory experiment. *Management Science* 57(4):611–627.
Herrmann, B., C. Thoni, and S. Gachter. 2008. Antisocial Punishment Across Societies. *Science* 319 (5868):1362–1367.
Korfiatis, N., E. Garcia-Bariocanal, and S. Sancherz-Alonso. 2012. Evaluating content quality and helpfulness of online product reviews: The interplay of review helpfulness vs. review content. *Electronic Commerce Research and Applications* 11(3):205–217.
Kumar, N., and I. Benbasat. 2006. Research note: the influence of recommendations and consumer reviews on evaluations of websites. *Information Systems Research* 17:425–439.
Ledyard, J.O. 1995. Public goods: a survey of experimental research. In *The handbook of experimental economics*, ed. A.E. Roth, J.H. Kagel, 111–194. Princeton: Princeton University Press.
Mas-Colell, J., M. Whinston, and J. Green. 1995. *Microeconomic theory*. Oxford: Oxford University Press.
Masclet, D., C. Noussair, S. Tucker, and M.-C. Villeval. 2003. Monetary and nonmonetary punishment in the voluntary contributions mechanism. *The American Economic Review* 93(1):366–380.
Mudambi, S.M., and D. Schuff. 2010. What makes a helpful online review? A study of customer reviews on Amazon. *Management Information Systems Quarterly* 3(4):185–200.
Park, D.-H., J. Lee, and I. Han. 2007. The effect of on-line consumer reviews on consumer purchasing intention: the moderating role of involvement. *International Journal of Electronic Commerce* 11(4):125–148.
Pavlou, P.A., H. Liang, and Y. Xue. 2007. Understanding and mitigating uncertainty in online exchange relationships: a principal-agent perspective. *Management Information Systems Quarterly* 31(1):105–136.

Reimers, I., and J. Waldfogel. 2020. *Digitization and pre-purchase information: the causal and welfare impacts of reviews and crowd ratings*. NBER working paper, Vol. 26776

Sahoo, N., C. Dellarocas, and S. Srinivasan. 2018. The impact of online product reviews on product returns. *Information Systems Research* 29(3):723–738.

Stephen, A.T., Y. Bart, C. Du Plessis, and D. Goncalves. 2012. *Does paying for online product reviews pay off? The effects of monetary incentives on consumers' product evaluations*. working paper.

Wang, J., A. Ghose, and P.G. Ipeirotis. 2012. *Bonus, disclosure and choice: what motivates the creation of high-quality paid reviews?* ICIS Proceedings 2012, ICIS 2012, Orlando (USA).

Weimann, J. 1994. Individual behaviour in a free riding experiment. *Journal of Public Economics* 54(2):185–200.

Zelmer, J. 2003. Linear public goods experiments: a meta-analysis. *Experimental Economics* 6(3):299–310.