Market Performance and Review System Informativeness

Philippos Louis*
Panos Markopoulos†
Georgios Voucharas‡
Dimitrios Xefteris§

Abstract

Review systems are the main tool used in online markets to cultivate trust among buyers and sellers. The proliferation of fake reviews on such systems could undermine users' confidence in the information they provide, ultimately hampering their ability to support the good functioning of such markets. We use a carefully designed lab experiment to study how the informativeness of such systems affects market efficiency and overall functioning. We find that a lack of confidence in reviews in either direction, i.e. positive, negative or both, is detrimental and leads to a lower volume of trade, lower quality, and a lower surplus generated as a result. Under some conditions, fixing the review system can restore trust in the market, although our results suggest that this is not the case following operation with a review system generating excessive positive reviews.

Keywords: reputation systems, online markets, trust, informativeness, lab experiment
JEL codes: D43, D83, L15

Funding: This work was co-funded by the European Union and the Republic of Cyprus through the Research and Innovation Foundation
(Project: EXCELLENCE/0421/0513)

*University of Cyprus, Department of Economics, Nicosia.
†University of Cyprus, Department of Business and Public Administration, Nicosia.
‡Liverpool Hope University, School of Business, Liverpool.
§University of Cyprus, Department of Economics, Nicosia.

1 Introduction

Central to the functioning and success of global e-commerce, projected to soon top $8 trillion in sales volumes (Statista, 2024), is the ubiquitous practice of allowing market participants to rate their experience with their counterparties. The "star-based" review system, where buyers observe a number of review stars, is probably the best-known example. Such reputation mechanisms provide a platform for users to evaluate the reliability and quality of sellers and their offerings, and help mitigate the inherent uncertainty associated with online transactions. They thus serve as the cornerstone of the effort of market operators to build trust among users and facilitate smooth market transactions. Indeed, many studies reveal that about nine out of ten consumers globally check reviews online before making purchases (Trustpilot, 2020, Uberall, 2021).

However, not all feedback and reputation systems are created equal. Detailed reviews, for instance, have been found to be more informative and useful than simple numeric ratings, suggesting that providing more granular feedback options could improve the informational content of feedback, thereby helping buyers make more informed decisions (Nosko and Tadelis, 2015). In general, the effectiveness of these systems can vary significantly, influenced by their design, the informational content they provide, how they are integrated into the marketplace's overall user experience, and how they supplement other sources of information, such as product and seller branding. In turn, these choices by the system's designer and the market context, affect the informativeness of the feedback mechanism, by influencing the behaviour of the users, including their willingness to submit a review, its informational content and even the degree to which users may use the system to artificially inflate or deflate sellers' ratings.

Indeed, deceptive, or "fake" reviews are a major threat to the informativeness of reputation systems, with consumer advocacy groups estimating that 30-40% of all online reviews can be considered "fake" (U.S. Public Interest Research Group (PIRG), 2023, Bloomberg, 2020), including fabricated, for-pay reviews, negative reviews targeting competitors, and reviews that do not disclose various quid pro quo arrangements. The number of fake reviews varies widely across markets and does not capture the full scale of the detrimental impact to markets, as it does not include the suppression of honest but unflattering reviews by sellers [1].

Despite the penalties imposed by many platforms on those caught posting fake reviews, including a recent proposal by the US Federal Trade Commission (FTC) for fines to businesses of up to $50,000 for each fake review, for each time a consumer sees it (Fowler 2023), the incentives for sellers to engage in such practices seem especially strong. Many studies showed that extra review stars translate to more sales (Luca, 2016) or that sellers with higher feedback ratings can com-

[1] this is especially prevalent in the market for smartphone apps. App developers regularly ask users if they "enjoy the app", asking those who answer positively to rate the app in some App marketplace, while asking only for private feedback from those who answer negatively.

mand significantly higher prices for the same goods, compared to those with lower or no feedback (Resnick et al., 2006).

At the same time, the detection of fake reviews appears to be getting more difficult, as research shows that fake reviews have become increasingly similar to genuine reviews over time so platforms can no longer easily use content to distinguish and filter fake reviews from genuine ones (Alma Economics, 2023). Recent advances in natural language generation have made convincing and difficult to detect fake reviews even easier and cheaper to produce.

Given the enormity of the role of reviews in global e-commerce, it is surprising that we know relatively little about the impact of fake reviews, and more generally, of how the informativeness of a feedback/reputation system is linked to the building of trust between market participants. Detecting such a correlation in the field is not easy. To measure how accurately a review left online corresponds to the reviewer's true buying experience requires knowing what this experience was, which is typically unobservable. The matter is further complicated by the fact that measures of informativeness typically provide only a partial order over information structures (Blackwell, 1951). For instance, it is not possible to order in terms of informativeness a system where half of the reviews are "fake positive" reviews with another where the opposite is true: half of the reviews are "fake negative". Still, one might imagine that these two systems will not be equally effective in promoting trust in the market.

The asymmetrical impact of positive versus negative feedback on consumer perceptions and seller behavior further complicates the dynamics of online marketplaces. Negative feedback, while less frequent, has a disproportionately large impact on consumer trust and decision-making processes. This phenomenon is partly because consumers often expect positive reviews and thus give more weight to negative information when it appears, a tendency that can significantly influence the reputation and success of sellers in the marketplace (Hu et al., 2009). The strategic manipulation of feedback, including the potential for inflated reputations due to a prevalence of positive feedback or the deterrent effect of negative feedback on seller behavior, underscores the critical role of feedback systems in maintaining quality standards and trust within online marketplaces (Dellarocas, 2006).

This paper aims to delve deeper into the nuanced role of feedback and reputation systems in online marketplaces, with a particular focus on the impact of fake reviews. We consider markets where buyers have no prior experience with any of the potential sellers and, for simplicity, we assume that no signals exist other than reviews about previous transactions and the current product price. Buyers can only rely on these when deciding whether or not to trust a specific seller. Our premise is that any such review system can be viewed as an information structure that provides buyers with a noisy signal about a specific seller's past behavior. The signal is inherently noisy as no two buyers can be expected to have exactly the same experience with the product or service, given diverse personal preferences, varying use cases, as well as random, external factors that may influence the performance of the products sold. Fake reviews further increase the noise and, depending on the relative prevalence of positive vs. negative fakes, introduce asymmetries in the information structure. We then ask the question: how does this information structure's informativeness and tendency to generate an excess of positive or negative reviews affect the performance of such a market? Furthermore, we examine whether market participants' experience in a market with a high volume of fake reviews can spillover and affect market performance (e.g. sales volume) in a different market where fake reviews are absent.

To overcome the difficulties inherent in studying these questions employing field data, we make use of a lab experiment. This presents its own set of challenges that we address. The most pressing issue, which determined the majority of our design choices, is that of recreating in the lab the one-shot nature of the buyer-seller interaction in online markets. In particular we need to ensure that when buyers in our experimental market need to decide if they want to buy from a seller, they can only condition this choice on the prices and reviews they observe, and not any prior interaction they had with the specific seller. We explain how we achieve this in Section 3.

In the experiment participants take up the role of either buyers or sellers in an "online market". In each market round sellers choose the quality of their good and the price. Buyers observe the prices, but not the quality offered by the different buyers. They also observe a seller's current review, which can either be 'good' or 'bad' and provides a noisy signal about the seller's choice of quality in the *previous* round. Hence, as is the case in real markets, reviews in our experiment provide no direct information regarding the utility that a buyer will get from a seller's product. They only provide information about the seller's past behavior. The accuracy of such information is our main treatment variable, which we manipulate to emulate the effect of fake reviews.

In the different treatments of the experiment participants start off in markets with different review informativeness for the first half (45 rounds). We have four treatments in total. One where informativeness is symmetrically high, i.e. if a seller chooses high (low) quality in one round they are very likely to receive a 'good' ('bad') review for the next round. In another treatment, informativeness is symmetrically low. In the other two treatments, informativeness is asymmetric. In the *bad*-informative treatment, when a seller chooses 'good' quality in one round, the review is almost certainly 'good' in the next round. If they choose low quality there is a substantial probability of the review being 'bad', but still it is more likely for it to be 'good'. Thus, overall there are mostly 'good' reviews but these are uninformative, as they could be the result of either choice in the previous round. On the other hand, a 'bad' review is very informative in this treatment. The reverse is true in the *good*-informative treatment.

In the second half (final 45 rounds), the reputation system's informativeness is symmetrically high in all four treatments. This allows us to examine how the experience with a specific system may influence behavior and market performance in a different market.

Our experimental findings confirm our intuition regarding the role of a reputation system's informativeness. In the symmetrically low informativeness treatment, the market performs poorly.

The overall surplus created is low, which is mainly due to a relatively low number of transactions. On the other hand, market performance is substantially better in the symmetrically high informativeness treatment, where the market generates a much higher surplus. The markets' performance in the asymmetric treatments is somewhere in between. Interestingly, though, we observe better performance in the *bad*-informative treatment compared to the *good*-informative one.

2 Related literature

The theoretical underpinning of this paper falls within the broader theme of "trust games", as described by Berg et al. (1995). That is a two-stage game with two players who are randomly paired, each of whom has an initial endowment. In the first stage, the first player decides how much of the endowment to send to the second player, and this amount is tripled by the experimenter. In the second stage, the second player receives the tripled amount and decides how much to pass back to the first player. While game theory predicts an equilibrium that is a corner solution (i.e., subjects decide to send none of their initial endowment to maximize their payoff), this prediction is not confirmed by the experiment itself. Interestingly, more recent studies also support these findings. [2] On average, participants send about half of their endowments to each other, revealing that trust and reciprocity are important behavioural aspects shaping economic decisions between subjects.[3].

The aforementioned "trust game" could be seen as a representation of an online marketplace, where the subjects act as buyers and sellers, illustrating the dynamics of trust. How does the presence of a reputation system influence the trust levels of buyers and sellers and consequently their decisions? The critical importance of feedback and reputation systems in building trust is not new. As of today, several iterations of the "trust game" have been employed, providing useful insights into how the informativeness of feedback and reputation systems as well as previous transaction history is linked to fostering trust between market participants. While the literature on reputation mechanisms spans several applications extending to fields such as economics, corporate environments and business networks, our focal point revolves around online marketplaces. [4] This Section summarizes and briefly describes key contributions to provide context and illustrate where our research fits into the existing body of literature.

Ba and Pavlou (2002) find that a better reputation system leads to higher levels of trust in an experimental setting. In another study, Pavlou and Gefen (2004) analyze questionnaire data from an online auction marketplace and show that effective institutional mechanisms positively affect buyers' trust. In an experimental study, Keser (2003) introduces a "trust game" a la Berg

[2]For a meta-analysis on trust games see Johnson and Mislin (2011).

[3]As Berg et al. (1995) show, with an initial endowment of 10$, on average, the first player sent $5.16 to the second one, which in turn, passed back $4.66.

[4]For a discussion on reputation literature, see also Abraham et al. (2016).

et al. (1995) that involves a reputation system. More specifically, at the end of the second stage, player 1 (i.e., the buyer), after receiving the amount from player 2 (i.e., the seller), is asked to evaluate his cooperation (i.e., negative, neutral or positive) which is then shared with her/him. The findings support that a reputation system is efficient and significantly increases the levels of trust and trustworthiness among subjects, leading to higher payoffs for buyers. In another experimental work, Chen et al. (2004) show that a relatively high informative reputation system which reveals past actions increases market performance and is more efficient compared to a relatively lower one. Contrary to earlier approaches, their experimental setup allows buyers to choose which sellers they wish to trade with (instead of being randomly paired) and buyers and sellers to decide whether they will fulfil their contracts. As argued by Bolton et al. (2004) reputation mechanisms play a vital role in building trust, especially in cases where repeated interaction between participants occurs. When buyers are aware that their past transaction behaviour is being observed, they tend to act honestly and fairly to maintain a good reputation (see for example Bracht and Feltovich, 2009). In addition, the presence of private information for building some reputation significantly enhances market performance (Huck et al., 2012). The findings of Abraham et al. (2016) suggest reputation mechanism improves trust, efficiency and profits even when the sharing of information comes at a cost compared to settings without information sharing. Unlike previous attempts, Dulleck et al. (2011) focus on a credence goods market in an experimental study and find that liability (i.e., providing sufficient quality goods) is a vital factor for improving efficiency while verifiability (i.e., observing and verifying the quality of goods provided) or reputation (i.e., identifying trading partners) building alone have little influence. Huck et al. (2016) examine the role of price competition and reputation in experience goods markets, where buyers are aware of the sellers' reputation and prices are set before the transaction. In the presence of price competition, sellers' motivation to build a reputation is significantly reduced. More recently, Zou et al. (2022) introduce an experimental game where sellers trade low- and high-quality products and buyers voluntarily evaluate their satisfaction with the transaction using a 5-point rating system. The rating system enhances market efficiency by improving the quality of the goods traded and increasing buyers' earnings. Tadelis (2016) offers an extensive review of the role of reputation systems in fostering trust and trade in online marketplaces.

Another strand of literature aims to investigate the role of genuine and deceptive reviews in influencing buyer behavior, trust levels towards sellers and market performance. How do reviews impact the purchasing decisions of buyers and their trust levels towards sellers? What is the impact of fake reviews on market performance? Do consumers perceive negative and positive reviews with equal significance?

Chevalier and Mayzlin (2006) utilize data from online booksellers and examine the impact of consumer reviews on sales. They show that a negative review has a stronger impact in reducing sales than a positive review in increasing sales. Ba and Pavlou (2002) show that negative feedback imposes a more negative impact on trust compared to positive ones. Interestingly, there is evidence that fake reviews increase the visibility of goods and thus can be beneficial for business in cases of positive fake reviews (Lappas et al., 2016). In the context of a field experiment, Akesson et al. (2023) find that fake reviews significantly affect consumer behavior, leading them to opt for products of lower quality resulting in significant welfare losses. While consumers usually identify products with negative reviews as low-quality ones, this is not the case for products with positive reviews which do not necessarily indicate high quality since even low-quality products can have some positive aspects (see also Herr et al., 1991 and Purnawirawan et al., 2015). Reimer and Benkenstein (2016) investigate the effect of online reviews on consumer behavior and find that positive reviews generally increase purchase intention more than negative ones. However, if a review is suspected to be untrustworthy, it will show a "boomerang effect" reducing purchase intention. The latter is also supported by Karabas et al. (2021). Additionally, the magnitude of the effect of the trustworthy reviews is greater than the effect of the untrustworthy ones. Maslowska et al. (2017) examine the effect of review valence on purchase intention and found that it is stronger when the volume of reviews is large or when high-price products are considered. Several important contributions have been made towards understanding the credibility and trustworthiness of reviews. Pooja and Upadhyaya (2022) provide an extensive and systematic review of studies on online review credibility. According to Kim et al. (2023), consumers react differently to credible compared to fake online reviews. Song et al. (2023) investigate how fake reviews shape consumers' purchase intentions for different types of goods. More specifically, a small number of false reviews enhances purchase intention for utilitarian products, while a higher number of fake reviews improves purchase intention for hedonic products.

Yet, several important questions remain unanswered. Our paper contributes to the growing body of theoretical and empirical literature on the system review informativeness on the spectrum of information asymmetry, where two parties (i.e., buyers and sellers) possess varying amounts of information (see also Akerlof, 1970) and concerns seller's characteristics on the potential risk of quality cheating (see also Mishra et al., 1998). Compared to the existing literature, our paper aims to implement a more effective simulation of an online marketplace through a controlled experiment in the laboratory where sellers interact with buyers in a one-shot game. Our experimental setup assures that buyers do not have prior knowledge of the sellers and do not expect future interactions with them. Importantly, buyers' decisions in our experimental market are driven by the prices and reviews they observe, while at the same time, they lack assurance that might come from a reputation effect. Our experimental setting offers both within- and between-subject variation in the review system specification. To pin down the implications of fake reviews on market performance and agents' decisions, we differentiate from the existing literature by not allowing buyers to determine the reviews of the sellers. Therefore, reviews are set externally and depend on each treatment which controls the informativeness of each review system. By doing this, we also address the challenge of the potential absence of reviews for sellers who have not completed any transactions.

3 Experimental design & Hypotheses

3.1 Main design and procedures

The experiment was carried out at the Laboratory for Experimental Economics at the University of Cyprus (UCY LExEcon). We recruited 144 participants from the lab's subject pool using ORSEE (Greiner, 2015) and conducted 10 experimental sessions in total: six with 12 and 4 with 18 participants per session. The experiment was computerised, designed and implemented in zTree (Fischbacher, 2007). All interactions were limited among participants in a given session and took place via computer terminals. At the beginning of each session, participants received written instructions that were read out aloud,[5] and then proceeded to the main experimental task. This consists of 90 rounds of trade in a market for a fictitious good, preceded by three trial rounds to familiarise participants with the environment and software interface. On average, each experimental session lasted for 75 minutes while the payoff per subject, including a five euro participation fee, ranged between 13.62 and 17.01 euros.

The experimental markets are made up of three buyers and three sellers. Participants' roles as buyers or sellers are assigned at the start of the experiment and remain fixed throughout. The composition of each market also remains fixed.[6] At the beginning of each round, each seller chooses the quality and price per unit of the good they offer. The quality can be either low or high, while the price ranges from integer values of 6 to 12. Next, buyers choose whether to buy one unit of the good from one of the sellers or an "outside option", which maintains a fixed price of 10 and is of high quality. When making their choice they observe the prices posted by each seller, but not the quality of the good offered. From round two onwards buyers also observe a Positive (green) or Negative (red) "review" signal for each seller. This depends stochastically on whether the specific seller chose high or low quality in the *previous* round. Therefore, this review does not reveal to buyers any information about the quality chosen by each seller in the current round. The way in which the review depends on the seller's previous choice of quality varies across treatments and is described in more detail below. The order in which the three sellers' offers appear on each buyer's screen is random and changes in each round. In this way, buyers cannot identify whether they have bought from the same seller before or not.

Each unit of a high-quality product that is sold incurs a cost of 3 points while each unit of a low-quality product that is sold incurs no cost. Thus, the difference between the price and the cost of the sold product determines the per unit profit of the sellers. Buyers start off the experiment with an endowment of 30 points. They value a high quality good at 12 points and a low quality good at zero. Thus, buying a high quality good results in a positive payoff. The payoff from

[5] The original instructions were written in Greek. We provide an English translation of these instructions in Appendix. The Greek version is accessible upon request.

[6] Each experimental session consisted of two or three independent markets.

choosing the "outside option" is always 2 points, i.e. the value of 12 points minus the price of 10 points.

At the end of each round, sellers are informed about their sales and profits. Before making their choice of quality and price in each round they observe the profile of prices, reviews and sales for all sellers from the previous round. They also observe the profile of reviews for all sellers in the current round. Buyers do not get informed in every round whether the product they purchased is of high or low quality (unless they have chosen the "outside option", which is always of high quality). Instead, every 3 rounds, they receive information about the total number of units bought in these rounds, their average payoff per round, their total payoff in the last 3 rounds, and their total points earned in all rounds.

At the end of the experimental session, the total number of points earned by each participant across all 90 rounds, including the initial 30 points endowment for buyers, is converted into euros at an exchange rate of one euro for every 35 points. Participants receive an additional five euro participation fee.[7]

3.2 Review system and treatments

The review system in the experimental market assigns each seller a review at the start of each round that depends on the quality chosen by the seller in the previous round. The review can be either *negative* or *positive* and appears to buyers along with the price chosen by each seller in *red* or *green* colour respectively. The dependence of the review on the previous round's quality choice is stochastic and captured by the pair of parameters $(\rho_{low}, \rho_{high})$, where ρ_j is the probability of the review being red, i.e. negative, given that the quality in the previous round was $j \in \{high, low\}$.

We used four different parameter specifications throughout the experiment. These are presented in the four panels of Table 1, where $(\rho_{low}, \rho_{high})$ are given in the first column of each panel.[8] While we allow for noise in all four specifications, the information structures differ substantially. The specific choice of parameters allows for a meaningful comparison between them, as we explain below.

Let one's prior belief about the probability that a seller chose bad quality in the previous round be denoted by $\beta_{t-1} = Pr(q_{t-1})$. After observing a negative review and by applying Bayes rule, the posterior probability becomes:

$$b_t(neg) = Pr(q_{t-1}|neg) = \frac{\rho_{low} \cdot \beta_{t-1}}{\rho_{low} \cdot \beta_{t-1} + \rho_{high} \cdot (1 - \beta_{t-1})},$$

[7] We provide screenshots capturing every phase of the laboratory experiment in Appendix xx.

[8] In the instructions and during the experiment, participants could see a table very similar to the panel corresponding to the specification used.

Table 1: Review system contingency tables used in the experimental treatments.

		review in t	
		negative	positive
quality	low	.9333	.0667
in $t-1$	high	.0667	.9333
$\kappa = 14$		$\lambda = 14$	

NF: No fake reviews

		review in t	
		negative	positive
quality	low	.35	.65
in $t-1$	high	.025	.975
$\kappa = 14$		$\lambda = 1.5$	

GF: Fake green (positive) reviews

		review in t	
		negative	positive
quality	low	.975	.025
in $t-1$	high	.65	.35
$\kappa = 1.5$		$\lambda = 14$	

RF: Fake red (negative) reviews

		review in t	
		negative	positive
quality	low	.60	.40
in $t-1$	high	.40	.60
$\kappa = 1.5$		$\lambda = 1.5$	

BF: Both fake reviews

which can also be written as:

$$b_t(neg) = \frac{\kappa \cdot \beta_{t-1}}{1 + (\kappa - 1) \cdot \beta_{t-1}},$$

where $\kappa = \frac{\rho_{low}}{\rho_{high}}$. This is one's *confidence in a negative review.* In two different review systems that have the same value of κ a buyer that has the same prior belief about a seller's quality in the previous round will end up with the same posterior beliefs after observing a negative signal. In a similar way we also define $\lambda = \frac{1-\rho_{high}}{1-\rho_{low}}$ as one's *confidence in a positive review.*[9]

The NF and BF specifications are both symmetric in terms of confidence in negative and positive reviews. In NF one can have high confidence in both reviews, while the opposite is true in BF. Specifications RF and GF are asymmetric. In RF one has a high confidence in positive reviews and low confidence in negative ones. The opposite is true in GF. The high confidence in positive and negative reviews, in RF and GF respectively, is equal to the confidence in both reviews one has in NF. At the same time, the low confidence in negative and positive reviews in RF and GF respectively, is equal to the confidence in both reviews one has in BF. Another thing to keep in mind is that for a given frequency of sellers choosing high quality under each review system specification, one expects to see an equal frequency of positive and negative reviews in the symmetric systems NF and BF, a high frequency of positive reviews in GF, and a high frequency of negative reviews in RF.

In the experiment we use both within- and between-subject variation in the review system specification. In particular, we have four between-subject treatments. In each treatment, markets start off with a different review system specification, which stays fixed for the first 45 rounds (phase I). After that, in all treatments, the review system switches to the NF specification and remains like that for the remaining 45 rounds (phase II). Participants are informed in the beginning that a

[9]See Weber (2010) for a detailed discussion of this approach to comparing information structures.

change in the review system *may* occur after round 45, but are not told what the new specification will be. They are informed about the new specification when the change occurs after round 45. In the treatment using NF in phase I, participants receive the exact same instructions and after round 45 are informed that the specification will remain the same for the remaining rounds. Table 2 provides a summary of the experimental design.

Table 2: Experimental design summary

Treatment	Phase I 45 rounds	Phase II 45 rounds	Markets	# of partcps	Obs.
Baseline	NF	NF	6	36	3240
All Fake	BF	NF	6	36	3240
Fake Positives	GF	NF	6	36	3240
Fake Negatives	RF	NF	6	36	3240

3.3 Discussion of the design

Perhaps the biggest challenge in designing the experiment was our aim to recreate the one-shot nature of the interaction between buyers and sellers, in a way that allows us to leverage the advantages of lab experiments. That means that buyers in our experiment should make their choices without being able to condition on their past interactions with a particular seller. Instead, they should rely only on the information presented to them in each round: seller's reviews and prices. Simply changing the position and labels of sellers randomly, as we do, is not sufficient. Persistent reviews would allow buyers to identify specific sellers even when put in different positions. To avoid this we limited the review memory to the minimum possible, i.e. one round. We also do not provide feedback to buyers in every round as this could allow them to make some inferences about sellers' identities based on the feedback and the review profile they observe.

While the main motivation for the experiment is the role of fake reviews, we do not allow buyers to affect the reviews of sellers. Reviews are set mechanically based on the specifications used in each treatment and phase. This is because our interest here lies in the effect of fake reviews on market performance. We therefore need to maintain tight control of the informativeness of the review system. Another advantage of doing so is that it avoids the issue of dealing with missing reviews when a seller has no sales.[10] This approach also helps to further ensure the one-shot nature of the interaction described above.

In all treatments, we switch the review system specification used in Phase I to the NF specification in Phase II. The purpose of this is to explore whether any loss of trust in the market

[10] Real reviewer systems often provide the number of reviews along with the review score. While it may be very interesting to understand the interplay between these two dimensions of a review system's information and market performance, this is left for future work.

that may arise in phase I –presumably due to fake reviews, persists even when a well-functioning review system is set in place. Ultimately we are interested in the potential for "contagion" between different markets, i.e. the abuse of a review system in one market leads to a failure of the system in a market with no widespread abuse. The setup here is a step in that direction.

3.4 Main Hypotheses

The theoretical predictions stated in Section 3 on the market trust game together with our formal findings allow us to propose a set of testable hypotheses.

To begin with, our focal point revolves around the way the informativeness of the RSIS affects market efficiency and the emergence of trust. Additionally, we are interested in investigating whether the erosion of trust under a "bad" RSIS can affect a review system's effectiveness in markets with a "good" RSIS.

Hypothesis 1: *Within a highly informative RSIS, the market will operate efficiently, generating a high surplus for all participants. If the RSIS is not very informative, the market will be inefficient with relatively few trades and a low average quality.*

We also expect informativeness conditional on the type of review (Green or Red) to matter.

Hypothesis 2: *Markets perform better with an RSIS where bad signals are rare but highly informative, compared to the case where bad signals are common but not very informative.*

4 Results

Table 3 summarises the experimental results concerning our main variables of interest. In this section we discuss these in detail, along with some additional analysis.

4.1 Market efficiency and surplus

In our experimental market, a unit of surplus is generated each time a high quality unit is traded between a buyer and a seller. On the flip side, surplus is lost in two ways. Either a buyer chooses the outside option instead of any of the three sellers' offers, or a buyer chooses the offer of a seller who chose to produce a low quality item. The first column of Table 3 shows the average surplus generated (as a percentage of the maximum possible surplus) in each phase of each treatment.

We first look at Phase 1. The Baseline (NF) treatment stands out here as the only one where markets generate a relatively high level of surplus. On average these markets generate 67% of the maximum possible surplus in the first 45 rounds. In contrast, in all three other treatments, the surplus generated is substantially lower and this difference is statistically significant. The average

Table 3: Market performance: aggregate results for each treatment by phase.

		Surplus % of max	Trade % freq.	Positive review % freq.	High quality % freq.	Price average	Buyers' payoff round avg	Sellers' profit round avg
Baseline	Phase 1	63	83	72	76	6.42	2.55	3.46
	Phase 2	64	81	71	76	6.27	2.94	3.17
All fake	Phase 1	38	54	50	65	6.99	1.72	2.60
	Phase 2	67	76	80	85	6.76	3.43	3.09
Green fake	Phase 1	41	65	82	63	6.59	1.37	3.02
	Phase 2	40	46	74	77	6.76	2.98	1.71
Red fake	Phase 1	39	61	24	66	6.43	1.46	2.78
	Phase 2	62	79	73	77	6.20	2.93	3.03

surplus generated in Phase 1 of these markets ranges from 38.5% of the maximum in RF to 41% in GF. Differences among these treatments are not statistically significant.

In Phase 2 the picture is different. In the Baseline treatment, where the specification of the Review system was NF in both phases, the average surplus remained at a similar high level as in Phase 1. Similar high levels in Phase 2 are also generated in the AF and RF treatments. There are no statistical differences across these three treatments in the average surplus generated in Phase 2. Treatment GF stands out by being the only one where the surplus remains very low in Phase 2, despite the change in the review system specification. The difference from the other three treatments is statistically significant.

The last two columns of Table 3 provide some indication about how the generated surplus is divided among the two sides of the market. Recall that buyers obtain a value of 12 points from a high quality unit, and zero otherwise. They also need to pay the price that can be between 6 and 12 points, which means that their payoff in a given round can range from -12 to 6 points. The outside option provides a high quality unit at a price equal to 10 points, which means they can guarantee themselves a payoff of 2 points per round. Given that the average price of traded units is below 7 points in all treatments, it is clear from the table that in all but the Baseline treatment, in Phase 1 of the experiment buyers on average make a loss when trading. The situation is better for buyers in Phase 2, where the average payoff is now close to or higher than 3 points. This is true even for the GF treatment where surplus is low in Phase 2. As we discuss in more detail below, this happens because although the volume of trade is low, reducing surplus, sellers choose high quality ore often. This means that while buyers buy from the market less frequently, they make a positive payoff on average when they do.

Sellers' profits in Phase 1 are relatively higher in the Baseline compared to the other treatments, although these differences are not statistically significant. Moving to Phase 2, sellers' profits increase significantly in AF and RF, reaching levels similar to those in the Baseline. Their profits fell slightly, but this difference is not statistically significant. The pattern reverses in treatment GF, where there is a substantial drop in sellers' profits.

4.2 Quality and volume of trade

Sellers' and buyers' choices jointly determine the possibility of surplus generation. Sellers do so by their choice of quality. Buyers on the other hand are ultimately the ones that determine whether trade occurs. Here we explore these choices to understand the role of each side of the market in the performance patterns we described above.

As can be seen in the 4th column of Table 3, in Phase 1 quality is relatively high in the Baseline and lower in all other treatments. This difference is statistically significant. In Phase 2 we find no differences across treatments, as quality reaches relatively high levels in all treatments, similar to the ones from Phase 1 in the Baseline. Figure 1 provides a more detailed picture. There one sees that the difference between the Baseline and the other treatments in Phase 1 can be attributed mostly to the sellers' choices in the first 15 rounds. In other words, faced with a high confidence review system specification sellers in the Baseline start off offering high quality straight away. In the other treatments with review system specifications where confidence for at least one type of review is low, sellers are more prone to offering low quality. These differences vanish in the later rounds of Phase 1 and treatments become indistinguishable in terms of quality offered in Phase 2 of all treatments.

The differences in quality in Phase 1 between the Baseline and the other treatments are in line with what we observed regarding the surplus generated. However, the sellers' choice of quality cannot explain the substantially lower surplus observed in Phase 2 of the GF treatment. We now turn to the buyers' side.

The 2nd column of Table 3 indicates that differences in buyers' choices in Phase 1 between the Baseline and the other treatments are in the same direction as those for the quality choices by sellers, and together, these give rise to the pattern observed for the surplus. That said, Figure 2 gives a more nuanced picture of the market dynamics in each treatment. In the first three rounds, there seems to be little to no differences across treatments, with buyers choosing to buy a unit from one of the sellers in about 80% or more of cases (78.9% in RF up to 90% in the Baseline). After that, in the Baseline the volume of trade remains relatively stable at this level. In AF and RF there is a rapid decline in the volume of trade. It stabilizes around 60% in RF, but drops even lower in AF, only to slightly recover towards the last third of Phase 1. In GF the decline follows a different pattern, progressing more gradually, but reaching the lowest level across all treatments in the final rounds of Phase 1. With the change of the review system specification in Phase 2, sales in AF and RF jump up and reach levels comparable to the Baseline. Only in GF do we see the volume of trade remaining at very low levels throughout Phase 2.

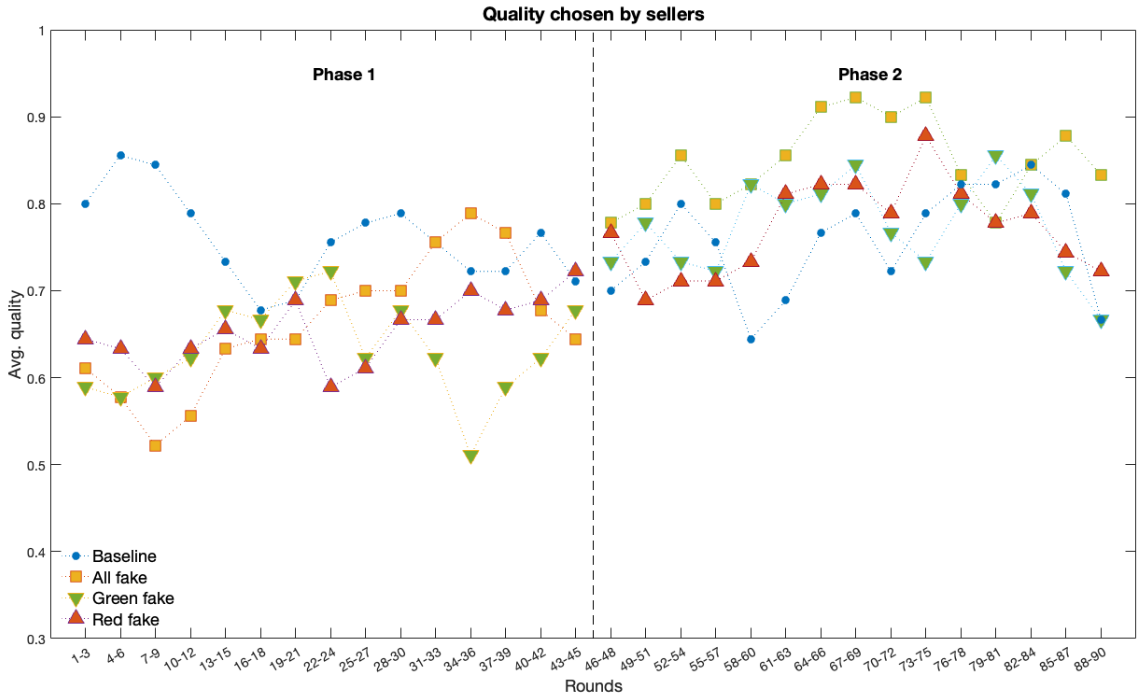

Figure 1: **Average quality chosen.** The points on the graph represent 3-round averages for the quality chosen by sellers, with a value of one representing high quality and zero for low quality. The vertical dashed line indicates the change from Phase 1 to Phase 2.

4.3 The role of reviews

The analysis above already provides clear evidence of a treatment effect. This means that The specification of the review system is critical for the market to be efficient. Here we take a closer look at how the review system affects buyers' choices.

Figure 3 shows how much of the trade happening in each treatment corresponds to sellers with positive or negative reviews. It already becomes clear from this that buyers trade mainly with sellers that have a positive review. The exception appears to be Phase 1 of the RF treatment, although one should note that in this treatment positive reviews are rare, due to the specific review specification (see the 3rd column in Table 3).

More generally, apart from a seller's review, a buyer's decision to engage with them can also depend on their price and the reviews and prices of the other sellers in the market. It is also clear from the previous analysis that the behavior of buyers can change substantially within a phase of the experiment as they adapt to market conditions. To better understand the role of reviews, we model a seller's sales volume as a function of their price, the difference of their price to the lowest price in the market, their review, the difference of their review to that of the other sellers, and the round in which trading takes place. We use Poisson regressions and the results are presented in Table 4.

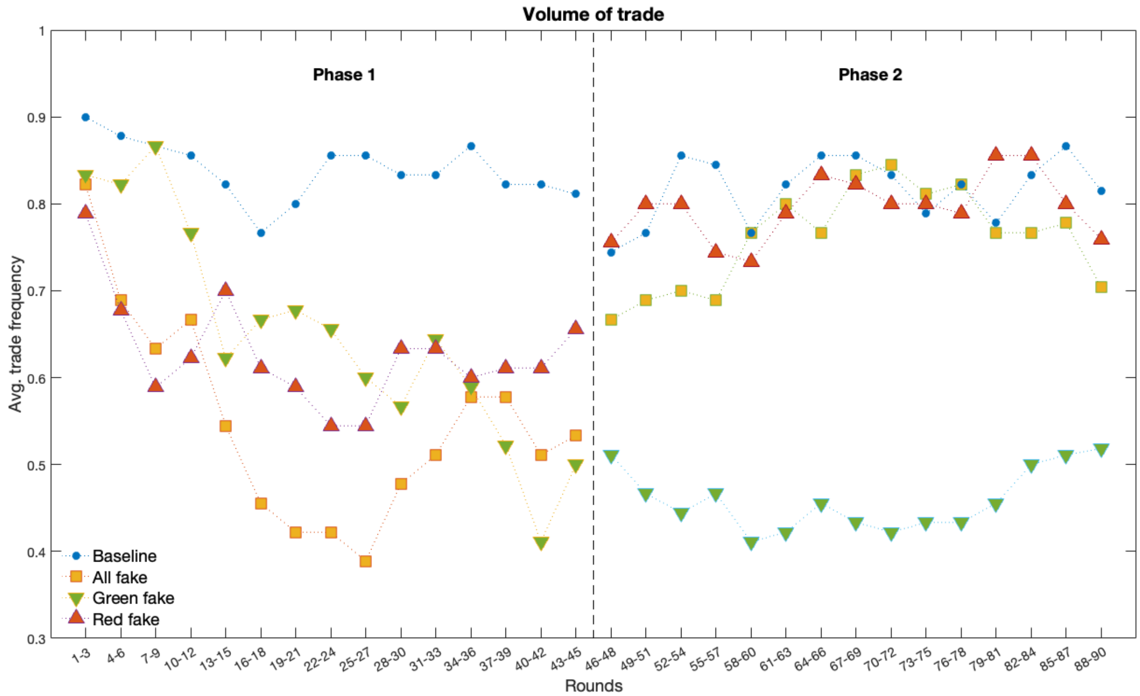

Figure 2: **Average volume of trade.** The points on the graph represent 3-round averages for the volume of trade as a percentage of the maximum possible. The vertical dashed line indicates the change from Phase 1 to Phase 2.

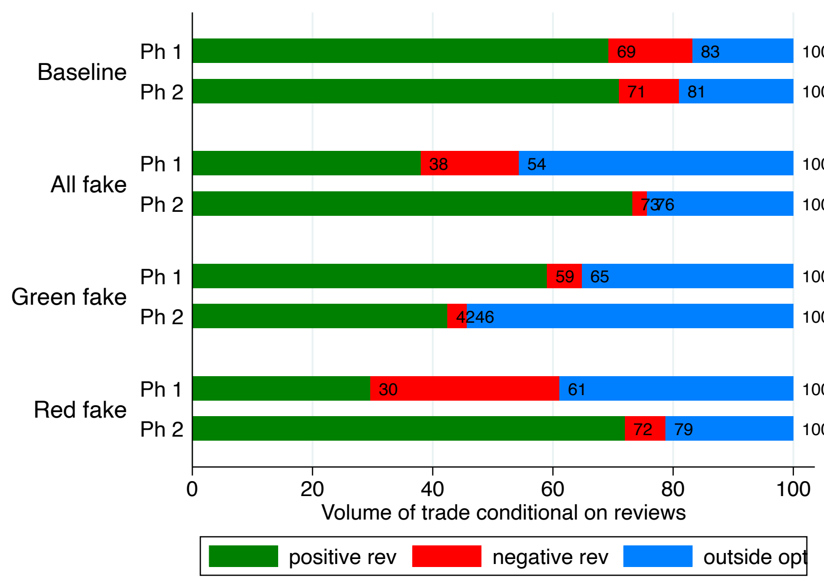

Figure 3: Volume of trade (%)

As expected, these regressions show that a seller's sales depend on the price, although the effect appears to be dependent mainly on the comparison of the price to the lowest available price, and not the absolute value. More interesting to our question, we find that reviews do seem to play the

dep var: *sales*	Phase 1				Phase 2			
	Baseline	AF	GF	RF	Baseline	AF	GF	RF
price	.070	-.253	-.090	-.149	-.041	-.029	-.418	.031
	(.273)	*(.047)*	*(.321)*	*(.387)*	*(.423)*	*(.590)*	*(.112)*	*(.895)*
price diff	-.238	-.039	-.207	-.269	-.293	-.356	-.071	-.443
	(.025)	*(.575)*	*(.091)*	*(.055)*	*(.022)*	*(.007)*	*(.555)*	*(.091)*
review	.147	.894	.389	.413	.445	1.636	1.351	.772
	(.131)	*(.006)*	*(.271)*	*(.137)*	*(.103)*	*(.004)*	*(.028)*	*(.013)*
review div	.553	.074	.635	.871	.633	.506	.090	.772
	(.004)	*(.748)*	*(.015)*	*(.042)*	*(.000)*	*(.063)*	*(.822)*	*(.008)*
round	-.001	-.008	-.015	-.002	.001	.001	-.006	-.001
	(.544)	*(.410)*	*(.036)*	*(.633)*	*(.647)*	*(.945)*	*(.239)*	*(.690)*
const	-.723	.786	.176	.345	-.388	-1.585	1.213	-1.030
	(.185)	*(.305)*	*(.763)*	*(.724)*	*(.308)*	*(.013)*	*(.430)*	*(.738)*
# obs	756	756	756	756	810	810	810	810
log Lik	-816.067	-665.165	-727.781	-683.235	-841.056	-808.430	-641.894	-830.444

Table 4: **Poisson regressions:** The coefficients in each column correspond to Poisson regressions of sales using data from the given phase of a specific treatment. For Phase 1 we exclude data from the first 3 rounds. The dependent variable in all cases is *sales* (a seller's sales in a given round, which varies from 0 to 3). The independent variables are: *price* (price chosen by seller), *price diff* (difference between seller's price and the lowest price of the market in that round,*review* (seller's current review, binary with 1 for positive review), *review div* (the difference between a seller's review and the average review of the other two sellers), *round*. The numbers in the parenthesis are *p-values* calculated using the wild bootstrap method, clustering std. errors at the group level.

critical role we anticipated. Depending on the treatment and the phase, reviews seem to affect sales either directly, or relative to the reviews of other sellers in the market.

We interpret the above results as evidence of the review system operating in an anticipated manner within markets. Namely, positive reviews positively influence a buyer's willingness to buy from a particular seller. Still, the preceding analysis also makes it clear that the effects of the review system can have consequences that go beyond an individual buyer's choices, affect both sides of the market and determine market performance more broadly. The specific way this happens depends on the details of the review system specification.

5 Discussion and conclusions

Our experiment is, to our knowledge, the first attempt to recreate in a controlled lab experiment the condition often found in online marketplaces, where established sellers have one-shot interactions with buyers. Given the nature of these interactions, there is typically no commitment device to ensure the quality of the good or service provided. At the same time buyers cannot rely on any past experience of their own with a particular seller, and at the same time do not anticipate interacting with them again in the future, which could provide some assurance through a reputation effect. Review systems have been the main tool used to solve this moral hazard problem.

The level of control afforded in the lab, allows us to focus entirely on the informational content of the review system. In particular, reviews in our experimental markets only provide information on a seller's past choices, but no information about their current quality. Under these conditions, the review system can work only if buyers adopt the social norm of "rewarding" positive reviews with their trust, and sellers anticipate such behavior and react positively to it. Such behavior from both buyers and sellers can lead to high levels of market efficiency, to be enjoyed by both sides of the market.

Our main finding is that such high efficiency is only possible when the review system is such that one can have high confidence in reviews corresponding to the implied past behavior. Any deviation from this standard leads to market failure, one way or another. Fake reviews are perhaps the most worrying trend observed in real online markets that could lead to the erosion of confidence in review systems. Our experiment constitutes the first experimental evidence of the detrimental effects fake reviews can have on market performance. Interestingly, these effects in our experiment are more severe than we anticipated, as they appear even with "semi-well-functioning" review systems like the ones in the GF or RF specifications.

A perhaps more alarming finding is that under certain conditions, the loss in trust between market participants can be irreparable. Fixing the review system and eliminating problematic factors such as fake reviews may not be enough to "fix the market". Given this result, one may speculate that the erosion of trust induced by fake reviews in one market can negatively affect the performance of other markets by association. If fake reviews have such negative externalities, tougher regulations or other measures to combat fake reviews might be called for. While at this stage we cannot rely on our experimental design to address the question of the existence of such externalities, our findings do suggest that it is in the realm of possibilities.

The role of fake reviews in online markets is a complex issue, many aspects of which we still do not understand fully. Acquiring such understanding requires an interdisciplinary approach, combining insights from computer science, psychology, marketing, information science and economics. The current paper is a step in that direction. It demonstrates the potential of carefully designed lab experiments to contribute to this exciting area of research and provides some new and surprising insight on the topic. A lot is still left to be done in future research.

References

Abraham, M., Grimm, V., Neeß, C., and Seebauer, M. (2016). Reputation formation in economic transactions. *Journal of Economic Behavior & Organization*, 121:1–14.

Akerlof, G. (1970). The market for" lemons": qualitative uncertainty and the market mechanism. *Quarterly Journal of Economics*, 89(3).

Akesson, J., Hahn, R. W., Metcalfe, R. D., and Monti-Nussbaum, M. (2023). The impact of fake reviews on demand and welfare. Technical report, National Bureau of Economic Research.

Alma Economics (2023). Fake reviews research. Accessed: 2024-06-02.

Ba, S. and Pavlou, P. A. (2002). Evidence of the effect of trust building technology in electronic markets: Price premiums and buyer behavior. *MIS quarterly*, pages 243–268.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and Economic Behavior*, 10(1):122–142.

Blackwell, D. (1951). Comparison of experiments. In *Proceedings of the second Berkeley symposium on mathematical statistics and probability*, volume 2, pages 93–103. University of California Press.

Bloomberg, I. L. (2020). Amazon fake reviews reach holiday season levels during pandemic.

Bolton, G. E., Katok, E., and Ockenfels, A. (2004). How effective are electronic reputation mechanisms? an experimental investigation. *Management science*, 50(11):1587–1602.

Bracht, J. and Feltovich, N. (2009). Whatever you say, your reputation precedes you: Observation and cheap talk in the trust game. *Journal of public economics*, 93(9-10):1036–1044.

Chen, K.-Y., Hogg, T., and Wozny, N. (2004). Experimental study of market reputation mechanisms. In *Proceedings of the 5th ACM conference on Electronic commerce*, pages 234–235.

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

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management science*, 52(10):1577–1593.

Dulleck, U., Kerschbamer, R., and Sutter, M. (2011). The economics of credence goods: An experiment on the role of liability, verifiability, reputation, and competition. *American Economic Review*, 101(2):526–555.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental economics*, 10(2):171–178.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, 1(1):114–125.

Herr, P. M., Kardes, F. R., and Kim, J. (1991). Effects of word-of-mouth and product-attribute information on persuasion: An accessibility-diagnosticity perspective. *Journal of consumer research*, 17(4):454–462.

Hu, N., Pavlou, P. A., and Zhang, J. J. (2009). Why do online product reviews have a j-shaped distribution? overcoming biases in online word-of-mouth communication. *Communications of the ACM*, 52(10):144–147.

Huck, S., Lünser, G. K., and Tyran, J.-R. (2012). Competition fosters trust. *Games and Economic Behavior*, 76(1):195–209.

Huck, S., Lünser, G. K., and Tyran, J.-R. (2016). Price competition and reputation in markets for experience goods: An experimental study. *The RAND Journal of Economics*, 47(1):99–117.

Johnson, N. D. and Mislin, A. A. (2011). Trust games: A meta-analysis. *Journal of economic psychology*, 32(5):865–889.

Karabas, I., Kareklas, I., Weber, T., and Muehling, D. D. (2021). The impact of review valence and awareness of deceptive practices on consumers' responses to online product ratings and reviews. *Journal of Marketing Communications*, 27(7):685–715.

Keser, C. (2003). Experimental games for the design of reputation management systems. *IBM Systems Journal*, 42(3):498–506.

Kim, J. M., cheon Park, K. K., and Mariani, M. M. (2023). Do online review readers react differently when exposed to credible versus fake online reviews? *Journal of Business Research*, 154:113377.

Lappas, T., Sabnis, G., and Valkanas, G. (2016). The impact of fake reviews on online visibility: A vulnerability assessment of the hotel industry. *Information Systems Research*, 27(4):940–961.

Luca, M. (2016). Reviews, reputation, and revenue: The case of yelp. com. *Com (March 15, 2016). Harvard Business School NOM Unit Working Paper*, (12-016).

Maslowska, E., Malthouse, E. C., and Viswanathan, V. (2017). Do customer reviews drive purchase decisions? the moderating roles of review exposure and price. *Decision Support Systems*, 98:1–9.

Mishra, D. P., Heide, J. B., and Cort, S. G. (1998). Information asymmetry and levels of agency relationships. *Journal of marketing Research*, 35(3):277–295.

Nosko, C. and Tadelis, S. (2015). The limits of reputation in platform markets: An empirical analysis and field experiment. Technical report, National Bureau of Economic Research.

Pavlou, P. A. and Gefen, D. (2004). Building effective online marketplaces with institution-based trust. *Information systems research*, 15(1):37–59.

Pooja, K. and Upadhyaya, P. (2022). What makes an online review credible? a systematic review of the literature and future research directions. *Management Review Quarterly*, pages 1–33.

Purnawirawan, N., Eisend, M., De Pelsmacker, P., and Dens, N. (2015). A meta-analytic investigation of the role of valence in online reviews. *Journal of Interactive Marketing*, 31(1):17–27.

Reimer, T. and Benkenstein, M. (2016). When good wom hurts and bad wom gains: The effect of untrustworthy online reviews. *Journal of Business Research*, 69(12):5993–6001.

Resnick, P., Zeckhauser, R., Swanson, J., and Lockwood, K. (2006). The value of reputation on ebay: A controlled experiment. *Experimental economics*, 9:79–101.

Song, Y., Wang, L., Zhang, Z., and Hikkerova, L. (2023). Do fake reviews promote consumers' purchase intention? *Journal of Business Research*, 164:113971.

Statista (2024). Retail e-commerce sales worldwide from 2014 to 2026. Accessed: 2024-05-20.

Tadelis, S. (2016). Reputation and feedback systems in online platform markets. *Annual Review of Economics*, 8:321–340.

Trustpilot (2020). The critical role of reviews in internet trust. Accessed: 2024-05-20.

Uberall (2021). Study: 67% of us consumers say fake online reviews a growing problem. Accessed: 2024-05-20.

U.S. Public Interest Research Group (PIRG) (2023). Technical report, U.S. Public Interest Research Group (PIRG).

Weber, T. A. (2010). Simple methods for evaluating and comparing binary experiments. *Theory and decision*, 69:257–288.

Zou, W., Wang, J., and Yan, J. (2022). Online markets and trust. *Journal of Economic Behavior & Organization*, 201:395–412.