Fake Reviews: Laboratory Evidence on Efficiency, Trust, and Trustworthiness *

Sharad KC†, Xiangyu Meng‡

November, 2026

Abstract

As consumers increasingly rely on online reviews to make decisions, sellers have an incentive to post fake reviews to build up their reputation and, consequently, sales revenue. Reviews partly solve the information asymmetry problem by signaling sellers' trustworthiness. Fake reviews, however, compromise this effect, distorting the online reputation market and leading truthful sellers to face unfair competition and buyers to make sub-optimal choices. However, the actual size of this impact is mostly unknown, because field data rarely distinguish fake reviews. Our paper uses a laboratory experiment in an experience goods market to measure the impact of fake reviews on market efficiency, trust, and trustworthiness, with review authenticity as a treatment. We found the opposite. Markets with fake reviews are more efficient than markets with verified reviews (49% against 31%). Fake review market does not change sellers' trustworthiness compared with verified reviews, and both fall short of the no-review baseline. The efficiency gain comes mostly from buyers' opt-out from high-stake transactions (46% in fake against 26% in verified). Our explanation for this odd results is that providing the right type of product in high-stakes costs more than the possible gain from raising ratings. More specifically, the implied reputation gain only covered 1-11% of the cost, 1% in fake and 11% in verified, which makes the trade totally not worth it. Fake reviews have a strong positive correlation with the ratings shown, and negative with trustworthiness.

Keywords: Fake reviews, trust, trustworthiness, efficiency

*We thank the faculty and staff of the Experimental Economics Lab at Georgia State University for their support. We also thank participants in the graduate student seminars and our classmates in the experimental economics course at Georgia State University for their helpful feedback. This research was supported by a Dissertation Fellowship grant from the Andrew Young School of Policy Studies at Georgia State University, which we gratefully acknowledge. We thank Enago (www.enago.com) for language editing of an earlier version of this manuscript.

†Zayed University, Georgia State University. sharad.kc@zu.ac.ae

‡New York University Abu Dhabi, Georgia State University. xiangyumeng@nyu.edu

1 Introduction

Reputation systems, as one solution to information asymmetry (Akerlof, 1978; Resnick et al., 2000), play a vital role in our daily lives. Buyers use ratings and reviews to make purchasing decisions, such as where to eat, shop, and stay. Sellers have incentives to build their reputation and consequently their revenue (Chevalier & Mayzlin, 2006; Luca, 2011); however, in the absence of a review verification process and a penalty system for fake reviews, sellers can manipulate reviews to increase sales (Luca & Zervas, 2016; Mayzlin et al., 2014).

When product quality is unobservable to buyers, markets rely on signaling mechanisms to mitigate information asymmetry, ranging from advertising (Nelson, 1970, 1974) to visual cues such as seller photos (Ert et al., 2016). Reputation systems go a step further. Rather than relying on the claims of a single seller, they aggregate feedback from many buyers into a single comparable measure of trustworthiness (Resnick et al., 2000).

The way buyers evaluate quality depends largely on whether a product is a search good (inspected before purchase), an experience good (evaluated only after use), or a credence good (relying on trust even after consumption) (Darby & Karni, 1973; Nelson, 1970). In e-commerce, shoppers cannot physically examine products, which increases uncertainty—especially for items where sensory features such as fit, material quality, or feel matter most (Dimoka et al., 2012; Peck & Childers, 2003). While online platforms make it easy to compare objective specifications, they still require consumers to judge sensory attributes without direct experience (L. R. Klein, 1998; Lal & Sarvary, 1999). As a result, buyers use user reviews and other seller-provided signals as indirect cues to assess product quality and reduce risk (Mavlanova et al., 2012; Mudambi & Schuff, 2010).

Digital platforms like Amazon, eBay, Uber, and Airbnb rely on reputation systems to cultivate trust. However, systematic distortions frequently compromise these mechanisms. For example, platforms often suffer from reciprocal reviewing, a strategic behavior in which buyers and sellers artificially trade positive ratings or withhold honest negative critiques to avoid retaliatory feedback (G. Bolton et al., 2013; Dellarocas & Wood, 2008), thus inflating ratings and diluting their overall informativeness. This deliberate external manipulation further exacerbates the loss of informational value, as review fraud has shifted from isolated self-promotion to highly organized and decentralized feedback markets (Luca & Zervas, 2016; Mayzlin et al., 2014). He et al. (2022) conducted an empirical study of these networks and showed how sellers exploit third-party social media groups to solicit manufactured five-star ratings in exchange for product rebates and commissions. Ultimately, these fabricated reviews distort search algorithms and artificially increase short-term demand, thus reducing consumer welfare by eroding the reliability of the platform's signaling mechanisms (He et al., 2022; Luca & Zervas, 2016).

In online marketplaces, review fraud manifests itself in various forms. For example, sellers can post positive reviews about themselves (Luca & Zervas, 2016; Mayzlin et al., 2014), submit negative reviews to damage competitors (Luca & Zervas, 2016), or buy post-sale positive reviews from verified buyers in exchange for gift cards (He et al., 2022). Platforms have implemented detection algorithms that detect behavioral clues, such as review duration and reviewer submission frequency; however, fake reviewers continuously adapt to these mechanisms (Luca & Zervas, 2016; Mukherjee et al., 2013). The fundamental limit of this problem comes from off-platform incentives that operate outside of the visibility of

marketplace algorithms (He et al., 2022). Academic investigation is therefore bounded by a lack of verified ground-truth datasets, forcing researchers to rely on heuristic proxy (Rayana & Akoglu, 2015). Consequently, distinguishing a fake review from an authentic one remains extremely difficult (He et al., 2022).

To bypass these observability limits, researchers used indirect empirical designs by exploiting institutional differences in verification standards between platforms (Mayzlin et al., 2014), analyzing datasets of reviews actively flagged by the platform's own filters (Luca & Zervas, 2016), or monitoring illicit review-buying groups on social media networks (He et al., 2022). Together, these studies show that sellers with weak reputations or close competitors are more likely to post or buy fake reviews, and that fake reviews increase ratings and sales only temporarily (He et al., 2022; Luca & Zervas, 2016; Mayzlin et al., 2014). However, these papers cannot show how fake reviews change the behavior of buyers and sellers because the quality that a seller actually delivers and the reason a buyer chooses to buy are not observed in field data.

Within the experimental domain, researchers manipulate the information available to buyers to observe how it influences a seller's incentive to mislead. This methodology is based on experimental evidence that shows that while asymmetric information and quality uncertainty depress trade (Plott & Wilde, 1980), reputation formation alone does not always restore market efficiency in traditional (Lynch et al., 1991) or digital experience-good markets (G. Bolton et al., 2004). Incorporating a modern rating structure, Solimine and Isaac (2023) experimentally shows that implementing a five-star aggregate rating system in the experience goods market reduces, but does not eliminate false advertising. Keser and Späth (2021) distort the rating system by only allowing a positive rating and found that it lowers trust. Only a few experiments let participants manipulate their own reputation: Lumeau et al. (2010) use an investment game in which both players can pay to edit the last rating they received, and Krügel and Paetzel (2024) use a repeated public good game in which participants rate each other's contributions and can improve the ratings they received, at no, low, and high cost. In both, manipulation inflates ratings, makes them less reliable, and lowers cooperation.

However, these two lines of experimental research have developed separately. Market studies examine reputation systems based on verified transactions (G. Bolton et al., 2004; Huck et al., 2016; Solimine & Isaac, 2023) or exogenous computer-generated noise (Louis et al., 2025), but do not allow sellers to post fake reviews. In contrast, rating-manipulation games allow participants to alter ratings (Krügel & Paetzel, 2024; Lumeau et al., 2010), but lack an actual market for goods. The market matters for one reason in particular. In an investment game or a public good game, every unit of trust or cooperation raises total surplus, so if fake ratings lower trust, they lower surplus. In an experience goods market, the buyer can refuse to buy, and when sellers' trustworthiness is low, a purchase lowers total surplus. If fake ratings make buyers trust less, efficiency can then increase. Our paper fills this gap by letting sellers post fake ratings, for themselves and for another seller, in an experience goods market in which buyers decide whether to buy and sellers decide what quality to deliver.

We build a sequential trust game in an experience goods market, in which quality is unknown to buyers at the time of purchase but fully revealed after consumption (Huck et al., 2016; Nelson, 1970). The buyer chooses whether to buy or take an outside option,

which measures trust, and the seller then chooses the quality to deliver, which measures trustworthiness. The price depends on the buyer's type, not on the quality delivered. We apply three review regimes as treatments: no review; verified review, in which only buyers who completed a transaction can rate the seller; and fake review, in which sellers can also post ratings for themselves and for another seller at no cost. Before trading begins, each session is assigned to one review regime, thus every buyer-seller pair within that session experiences the same review environment. Because treatment varies only between nine sessions, three No Review, four Verified Review, and two Fake Review, we report two analyses of each contrast in Section 4, one with sessions and one with participants as the unit of observation. We also state a reputational incentive constraint that shows when a verified review system can sustain trustworthy behavior and why this condition weakens once sellers can post ratings at no cost.

We find that fake review treatment shows higher efficiency than verified review treatment in transactions with type A buyers, who need the high-quality good (48.7% versus 30.5%). This is not because sellers are more trustworthy: trustworthiness toward type A buyers is 14.7% and 15.4%. Sellers remain trustworthy toward type B buyers, who need only the low-quality good, since trustworthiness costs them nothing in that transaction, and they are untrustworthy toward type A buyers, where it is costly. Buyer behavior drives efficiency gain. Under fake reviews, type A buyers show less trust (26.3% versus 45.9%), and because type A trust lowers the total surplus whenever trustworthiness is below 5/6, as is the case in all sessions, lower trust raises efficiency. The difference in trust is significant with participants as units ($p = 0.030$) but not with sessions as units ($p = 0.200$), most of it comes from one of the two Fake Review sessions, and the difference in efficiency is not significant between sessions ($p = 0.133$).

We then ask why the reviews did not make the sellers more trustworthy. Using variation within sessions, we find that buyers' trust responds to displayed ratings, but a trustworthy transaction moves a seller's average rating by only 0.23 rating points under verified reviews and 0.03 under fake reviews, so the implied reputation gain is roughly 11% and 1% of its 6-point cost. Fake reviews therefore had little room to lower trustworthiness. Under fake reviews, the displayed rating mostly reflects sellers' own posts, and sellers who post more fake ratings are less trustworthy toward type A buyers. The Appendix compares both review systems with a no-review baseline, where trustworthiness toward type A buyers is surprisingly higher (42% against about 15%), but the difference is not statistically robust.

The remainder of the paper proceeds as follows. Section 2 presents our experimental design, Section 3 provides the reputation constraints and our hypotheses, and Section 4 presents the results. Finally, Section 5 discusses them.

2 The Experiment

2.1 Game Design

The experimental framework reflects the informational structure of experience-goods markets on online platforms, such as hotel or restaurant booking, where product quality remains unobservable until the transaction is finalized and the good is "experienced" (Darby & Karni, 1973; Nelson, 1970). The marketplace consists of two primary parties, buyers and sellers.

Buyers are divided into two types; type A (High) and type B (Low)—which correspond to two product qualities: Good A (High) and Good B (Low). Type A buyers can only gain utility from high-quality Good A. While type B buyers can be satisfied by Good A, the low-quality Good B is enough for their needs and costs the seller less.

A defining feature of the game is that the price paid by the buyer is exogenous to the realized quality; instead, it is strictly determined by the buyer's assigned type. This mirrors real-world service transactions—such as a hotel reservation, where a consumer selects and pays for a category of service upfront. The seller then retains the discretion to provide high- or low-quality service. This structure introduces a significant moral hazard, as the seller's revenue is decoupled from the quality of the good actually delivered (B. Klein & Leffler, 1981; Shapiro, 1983).

As illustrated in Figure 1, the stage game proceeds through a structured sequence of chronological events following a trust game design(Berg et al., 1995). Initially, the seller establishes prices p_a and p_b for both high-type and low-type goods, which are associated with production costs c_a and c_b such that $p_a > p_b$ and $c_a > c_b$. Subsequently, "Nature" assigns a buyer type (A or B), a realization that is immediately disclosed to both the buyer and the seller. This specific disclosure ensures that the buyer is aware of their requirements prior to the transaction, a defining characteristic for experience good markets. Upon observing the posted prices for both qualities and their assigned type, the buyer chooses to either enter the market ("In") or "Opt Out"; choosing the latter concludes the game for that round. If the buyer enters, the seller determines the quality of the good provided, and the buyer pays for the good.

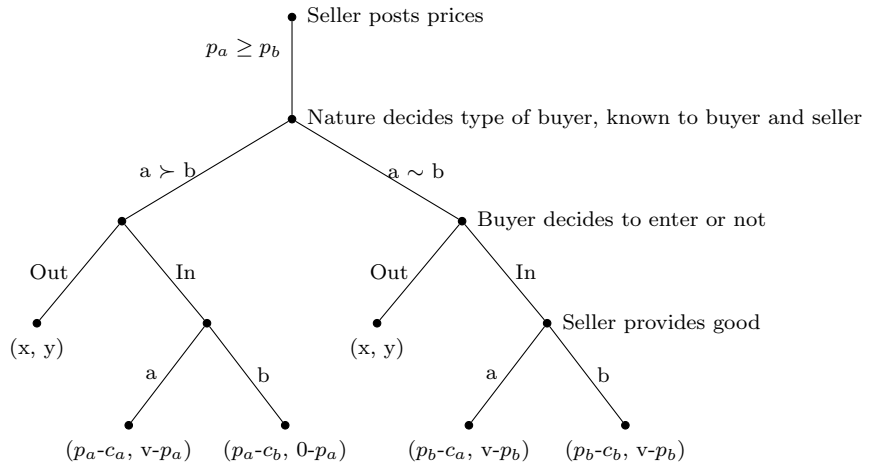


Figure 1 Experience Good Game Tree

The terminal nodes in Figure 1 illustrate the payoffs for both agents in the format (seller's payoff, buyer's payoff). The **seller's payoff** is defined as the price charged (p) minus the cost of the good provided (c). The **buyer's payoff** is determined by the alignment between the delivered quality and the buyer's assigned type: a buyer realizes a valuation v if the quality meets his/her requirements and pays a price determined by his/her type. Under this specification, a type b buyer always receives v regardless of seller's choice of quality.

In contrast, a type a buyer only receives v if the seller provides good a; if provided with a low-quality good b, the high-type buyer's valuation is 0, resulting in a net payoff of $-p$.

2.2 Treatments

The final stage of each round of the game, not shown in Figure 1, is dictated by the specific review treatment protocol.

In the baseline game (No Review), no reputation mechanism exists and the game ends after the transaction. Buyers and sellers possess only private information regarding their previous transactions, with no external feedback or ratings.

In the Verified Review treatment, following the reputation experimental designs by G. Bolton et al. (2004) and Solimine and Isaac (2023), rating is restricted to buyers who have successfully transacted with their matched seller. The buyer provides an integer rating between 1 and 5 after the transaction in each round, with 1 being the lowest and 5 the highest. These verified ratings serve as evidence of genuine interaction, intended to foster transparency and reliability within the marketplace. From the second round, the system informs the buyer about the arithmetic mean of these ratings, enabling them to evaluate a potential trading partner's reputation while deciding whether to enter the market.

The Fake Review treatment allows the seller to provide a rating for himself/herself and for another seller at the end of each round, regardless of the transaction occurrence. This system facilitates the inclusion of non-transactional feedback, allowing strategic manipulation and dissemination of deceptive information (Luca & Zervas, 2016; Mayzlin et al., 2014). The system also reveals aggregate seller rating to buyers after round 2, which now incorporates both authentic buyer feedback and non-authentic seller ratings.

Table 1 Treatments and sessions

Treatment	Sessions	Subjects	Subjects per session	Pair-round observations
No Review	3	38	10–18	304
Verified Review	4	40	8–12	320
Fake Review	2	40	18–22	320
Total	9	118	8–22	944

Table 1 provides the treatment table with number of subjects in each treatment[1]. Treatments are session based. The subjects are evenly divided between treatments; 118 undergraduate students participated in nine sessions (three No Review, four Verified Review, and two Fake Review), generating 944 pairs of observations over 16 paid rounds. All 118 participants are included in every analysis.

2.3 Matching Between Subjects

The role of a participant is assigned randomly during seating, and it does not change throughout the session. The matching between buyers and sellers is governed by a randomized

[1]See Table A.4 in the Appendix for incidence of fake review for each round of the game.

mechanism in which participants are paired anew at the end of every round. This random matching protocol ensures that neither buyers nor sellers retain any knowledge or identification of their previous counterparts, preventing the formation of reputations outside of the experimental treatments, minimizing strategic maneuvers or biases that might otherwise stem from repeated interactions with the same partner (Berg et al., 1995; G. Bolton et al., 2004). Matching takes place within a session: the participants in a session form the pool from which pairs are drawn each round, and all of them share the same review environment.

Under this framework and the no communication outside of the experiment system, buyers are granted no insight into the identities or characteristics of sellers they have previously engaged with; likewise, sellers remain unaware of their historical interactions with specific buyers. This anonymity is critical for maintaining the integrity of the experimental design and emulating real-world market conditions, where transactions often take place between strangers.

2.4 Experimental Procedure

The experimental session has eighteen rounds in total, organized into two phases. The preliminary phase includes two practice rounds, providing participants with the opportunity to familiarize themselves with the digital interface and the game-mechanics. The second round shows the average rating if available for the review treatments. This introductory stage also allows subjects to ask questions and seek clarification on the rules, ensuring a comprehensive understanding of the task before transitioning to the paid rounds.

The subsequent phase includes sixteen paid rounds during which participant decisions result in tangible economic consequences and rewards. This structural division between practice and paid rounds serves as both a pedagogical tool for participants and an analytical tool for researchers, as it helps distinguish initial exploratory behavior from the more deliberate and strategic decision-making that occurs in the later rounds.

The experiment was designed and conducted in the Georgia State University Experimental Economics Center (ExCEN) between April 4, 2022 and April 14, 2022, using oTree version 2.1.28(Chen et al., 2016). The study was determined exempt (Category 2, 3) by the Institutional Review Board of Georgia State University (protocol H22137, determination date October 5, 2021), and participants signed an informed consent form before the session.

Upon arrival at the lab, participants were briefed on the study's details and provided with informed consent, after which they were randomly assigned to computer terminals. Once all participants were seated, they received printed instructions and asked to sign the consent form[2]. To help participants understand the experimental task, there were two practice rounds of the game. The experimenter then answered any clarification questions that the participants had. Information regarding the distribution across treatment conditions was not disclosed beforehand, and participants were prohibited from communicating with each other during the experiment. After the end of the experiment, all participants answered a questionnaire asking for their demographic information. Participants did not receive monetary incentive for the questionnaire. Table A.2 reports the gender, age, and race of the 118 participants by treatment, taken from the post-experiment survey: 57% to 68% are female,

[2]Subject instructions for all treatments, the consent form, and the questionnaire are in the Online Appendix.

the mean age is about 20, and most participants are Black (53% to 62%) or Asian (17% to 33%) in every treatment. All participants received a show up fee of 5 US dollars. On average, the subjects earned a total of 20 US dollars.

3 Reputational Constraint and Hypotheses

This section develops the theoretical predictions of the game. We first derive the payoff functions for buyers, sellers, and the market, then use comparative statics to characterize the equilibrium of the no-review baseline and discuss how each review system changes sellers' incentives. While not solving the repeated game with ratings, we discuss a reputational constraint for interpretation of the Verified and Fake Review treatments. The resulting predictions form the hypotheses that are tested in Section 4 and Appendix 2.

3.1 Payoff Functions

The game tree in Figure 1 shows the potential payoffs for buyers and sellers in different scenarios. It is a trust game with a quality choice: the buyer's entry is the trust decision, and the seller's delivery of the type-appropriate good is the trustworthy response (Berg et al., 1995). The payoff for buyers, sellers, and the market as a whole below are functions of the probability of trust (β), trustworthiness (γ), buyer type (α), and posted prices (p_a, p_b).

$$\begin{aligned}
\text{Buyer} &= (1-\beta)y + \beta\big\{\alpha\big[\gamma(v-p_a) + (1-\gamma)(-p_a)\big] + (1-\alpha)(v-p_b)\big\} \\
&= (1-\beta)y + \beta\big[(1-\alpha+\alpha\gamma)v - \alpha p_a - (1-\alpha)p_b\big] \\
\text{Seller} &= (1-\beta)x + \beta\big\{\alpha\big[\gamma(p_a-c_a) + (1-\gamma)(p_a-c_b)\big] + (1-\alpha)\big[\gamma(p_b-c_b) + (1-\gamma)(p_b-c_a)\big]\big\} \\
&= (1-\beta)x + \beta\big[\alpha p_a + (1-\alpha)p_b + (c_a-c_b)(\alpha+\gamma-2\alpha\gamma) - c_a\big] \\
\text{Total} &= (1-\beta)(x+y) + \beta\big[(1-\alpha+\alpha\gamma)v + (c_a-c_b)(\alpha+\gamma-2\alpha\gamma) - c_a\big]
\end{aligned}$$

$\alpha \in \{0,1\}$ is the buyer's type: $\alpha = 1$ is type a (high type) and $\alpha = 0$ is type b (low type). β is the probability that the buyer chooses *In* (trust); $1-\beta$ is the probability of *Out*. γ is the probability that the seller provides the type-appropriate good (trustworthiness); $1-\gamma$ is the probability of a mismatch. x and y are the outside option payoffs for the seller and buyer, respectively.

The partial derivatives of the payoffs with respect to β, γ, p_a, and p_b are in Table 2. Prices affect only the distribution of surplus/payoffs between buyer and seller, not total surplus. Substituting the experimental parameters $x = y = 3$, $c_a = 10$, $c_b = 4$, $v = 18$ gives us the simplified partial derivatives in functions of α, β, γ, p_a, and p_b (Table 3).

3.2 Treatments

3.2.1 Baseline

In the baseline treatment, buyers and sellers are randomly matched each round with no public history of past behavior. Since no reputation signal exists and subjects are anonymous, the

	Buyer	Seller	Total
$\partial/\partial\beta$	$-y+(1-\alpha+\alpha\gamma)v$	$-x+\alpha p_a+(1-\alpha)p_b$	$-(x+y)+(1-\alpha+\alpha\gamma)v$
	$-\alpha p_a-(1-\alpha)p_b$	$+(c_a-c_b)(\alpha+\gamma-2\alpha\gamma)-c_a$	$+(c_a-c_b)(\alpha+\gamma-2\alpha\gamma)-c_a$
$\partial/\partial\gamma$	$\beta\alpha v$	$\beta(c_a-c_b)(1-2\alpha)$	$\beta\big[\alpha v+(c_a-c_b)(1-2\alpha)\big]$
$\partial/\partial p_a$	$-\alpha\beta$	$\alpha\beta$	0
$\partial/\partial p_b$	$-(1-\alpha)\beta$	$(1-\alpha)\beta$	0

Table 2 Partial derivatives of payoffs with respect to β, γ, p_a, and p_b

	Buyer	Seller	Total
$\partial/\partial\beta$	$15-18\alpha(1-\gamma)$	$\alpha p_a+(1-\alpha)p_b$	$6\gamma+6\alpha\gamma-12\alpha+2$
	$-\alpha p_a-(1-\alpha)p_b$	$+6(\alpha+\gamma-2\alpha\gamma)-13$	
$\partial/\partial\gamma$	$18\beta\alpha$	$6\beta(1-2\alpha)$	$\beta(6+6\alpha)$
$\partial/\partial p_a$	$-\alpha\beta$	$\alpha\beta$	0
$\partial/\partial p_b$	$-(1-\alpha)\beta$	$(1-\alpha)\beta$	0

Table 3 Partial derivatives of payoffs with respect to β, γ, p_a, and p_b with experimental parameters plugged in

stage-game equilibrium is predicted in every round: with a known end and a unique stage equilibrium, repetition does not change the prediction for selfish, rational subjects.

Type a buyer ($\alpha = 1$). The seller's payoff is decreasing in trustworthiness: $\frac{\partial \text{Seller}}{\partial \gamma} = -6\beta \leq 0$. Providing good a costs 6 more than providing good b, with no offsetting future benefit. The seller's dominant strategy is therefore $\gamma = 0$. Anticipating this, the type a buyer's expected payoff from entering is $-p_a$, indicating the dominant strategy is *Out* ($\beta = 0$). The unique subgame perfect equilibrium for type a is: buyer chooses *Out*; seller provides good b if entry were to occur.

Type b buyer ($\alpha = 0$). The seller's payoff is strictly increasing in trustworthiness: $\frac{\partial \text{Seller}}{\partial \gamma} = 6\beta > 0$. Since good b is also the cheaper option for the seller, the dominant strategy is being trustworthy, i.e. $\gamma = 1$. type b buyer enters whenever the participation constraint holds: $v - p_b \geq y$, i.e. $p_b \leq 15$. The sub-game perfect equilibrium for type b is: buyer chooses *In* if $p_b \leq 15$; seller provides good b.

Conflict of Incentives. The derivative $\frac{\partial \text{Seller}}{\partial \gamma} = 6\beta(1-2\alpha)$ reveals the structural conflict at the heart of this market. For type a buyers, the seller's incentive directly opposes market efficiency: trustworthiness costs 6 with no direct compensating benefit. For type b buyers, the seller's interests align with the buyer and with total welfare. This asymmetry — trustworthiness is costly for the seller when it benefits mostly the buyer — is the central market failure that review systems attempt to resolve.

Total payoff increases with buyer entry when $\frac{\partial \text{Total}}{\partial \beta} > 0$. For type a buyers ($\alpha = 1$), this requires $12\gamma - 10 > 0$, i.e. $\gamma > 5/6$. Since the seller's dominant strategy in the baseline is $\gamma = 0$, type a participation reduces total payoff, and the outcome of mutual *Out* is in fact market efficient. A type a buyer's own participation constraint is even stricter: entering

pays only if $18\gamma - p_a \geq 3$, i.e. $\gamma \geq (p_a + 3)/18$. γ needs to be no less than 0.94 at an average posted price of 14, so exit is the payoff-improving response whenever sellers deliver less often than that.

If subjects care only about their own payoff, sellers never deliver the costly good in the baseline. If some sellers also care about the buyer or about honesty, some will deliver even though it costs them(Berg et al., 1995). Buyers cannot tell which sellers are which.

3.2.2 Verified Review

The verified review system creates a public signal linking the seller's past trustworthiness to future buyer participation. A seller's action in the current round now affects the probability that future buyers choose *In*, transforming the one-shot game into a market-level repeated game. For type a buyers, the seller's short-run incentive for being trustworthy remains negative as $\frac{\partial \text{Seller}}{\partial \gamma} = -6\beta < 0$. Following the incentive constraints of the theory of reputation by B. Klein and Leffler (1981) and Shapiro (1983), a seller will voluntarily provide Good a only if the long-run reputation gain offsets the marginal cost of $c_a - c_b$, hence the reputational incentive constraint:

$$\sum_{k=1}^{T-t} \delta^k \, \Delta\pi_{t+k} \;\geq\; c_a - c_b = 6,$$

where $T = 16$ is the last paid round, t is the current round, and $\Delta\pi_{t+k}$ is the increase in the seller's expected profit in round $t + k$ from attracting a higher entry rate β. Because all earnings are paid together at the end of the session, rounds are not discounted for payment delay, so we set $\delta = 1$ and the sum runs over the rounds that remain. In the last round, the sum is empty and the constraint fails. Buyers can respond to a seller's record because they do not know the seller's trustworthiness. A good record makes buyers more willing to enter, which gives a selfish seller a reason to deliver early to keep that belief alive (Kreps et al., 1982).

The constraint is a reference point, not a prediction that we derive. We do not solve for how buyers' beliefs would respond to ratings, so we measure that response from the data. Theory also gives no guarantee that reputation sustains honesty here, because the game has a known end and sellers are matched with a different buyer each round. If the constraint holds, reviews can raise type a trustworthiness above the baseline. If it does not hold, the dominant strategy is to provide good b regardless of the review system, and the market outcome reverts to the baseline. The constraint parallels with the one in Dellarocas (2005), where a seller cooperates only if the discounted loss from a higher probability of a negative rating outweighs the cost of effort.

The constraint tells us when reviews can change this benchmark. If sellers expect future buyers to reward a good record enough to cover the 6-point cost, the standard theory of reputation predicts the following.

H1: Verified reviews raise type a trustworthiness relative to the no-review baseline. type b trustworthiness is unaffected.

Two aspects of the environment may make it harder to satisfy the constraint. First, the displayed rating averages over all of a seller's past ratings, which pool buyers of both types, so a single delivery changes it by roughly one over the number of ratings behind

it. type b buyer can almost always expect good b from sellers and then give them higher ratings. Second, a rating system can crowd-out honesty that sellers would otherwise provide for intrinsic reasons(Bowles & Polanía-Reyes, 2012; Frey & Oberholzer-Gee, 1997). We measure the first in Section 4 and treat the second as a conjecture that the experiment does not separate.

3.2.3 Fake Review

In the fake review treatment, sellers can post self-ratings that are not tied to their actual behavior. The reputation signal then reflects the true trustworthiness only in part, so buyers decide on their own experience and a contaminated average rating. This weakens the reputational incentive constraint: a seller can raise their ratings without paying the cost $c_a - c_b$. For a selfish seller, the long-run incentive to be trustworthy is no greater than verified review (H2). Knowing that ratings are a mixture of fake and real ratings, buyers can discount ratings, trust less, and stay out (H3). Since type a entry reduces total surplus whenever $\gamma \leq 5/6$, less type a participation leads to higher type a efficiency.

H2: type a trustworthiness is lower under Fake Review than under Verified Review. type b trustworthiness is not significantly different in the two review treatments.

H3: type a trust is lower under Fake Review than under Verified Review. type b trust is not significantly different in the two review treatments.

These hypotheses outline the expected relationships between ratings, trust levels, and trustworthiness under different treatment conditions.

Regarding who would post more fake ratings, theory is not clear. In Mayzlin (2006), firms with inferior products promote more, because honest posters already publicize the higher quality product. In Dellarocas (2006), it depends on how a firm's revenue rises with its perceived quality. We are testing, when sellers can provide different quality products, whether sellers who post more fake ratings deliver lower quality.

A seller can also post a rating for one other seller. Because a seller's payoff depends only on its own displayed rating, such postings give the poster no payoff gain, and the seller who receives one cannot tell who posted it. They show whether sellers use a costless, anonymous action to raise or lower another seller's rating.

4 Results

4.1 Evidence and Inference

We report two analysis of each contrast between Fake and Verified Review. Following common practices in experiments on feedback systems (G. Bolton et al., 2013; G. E. Bolton et al., 2019; Gazzale & Khopkar, 2011), our first uses the session as the unit, summarized by each session's mean outcomes for each buyer type (Figure 2 and Table A.1), with exact permutation tests on session means (Table 4). Because treatment was assigned to whole sessions, it assumes nothing about dependence within sessions. The tests enumerate every reassignment of the six sessions to the two treatments, so a p-value cannot fall below one over the number of assignments, which is 0.067 for 15 assignments, however large the difference.

The second analyses uses the participant as the unit, with the participant's mean over the relevant rounds and the Mann–Whitney tests (Table 5). It assumes that no shock is shared by all sessions of a treatment[3]. Participants in a session are no more alike than participants in different sessions of the same treatment (intra-session correlation 0.00 to 0.02[4]), but that comparison is within a treatment and cannot detect a shock shared by all sessions of a treatment. Neither test treats rounds as independent, since the same participant decides in all 16: each uses one number per session or per participant. As a supplement, Appendix Table A.3 reports linear probability models that use every participant-round as an observation. Their standard errors are clustered by participant, which allows a participant's rounds to be correlated, and they rest on the same assumption as the participant-level tests.

Comparisons with the No Review sessions, including a contrast that pools the six review sessions, are in Appendix 2 (Tables A.5 and A.7). No session-level p-value in either table is below 0.05. We make no correction for multiple testing; with a Bonferroni correction over the four participant-level tests of Fake versus Verified Review, none would be significant.

Table 4 reports on three main outcomes: trustworthiness, trust, and efficiency. Trustworthiness is defined as the seller providing the buyer with the desired quality good, i.e. good a for type a buyer and good b for type b buyer. Trust is measured by the buyer choosing to enter the market and transact with the matched seller. Market efficiency is calculated based on total payoffs, defined as the sum of realized payoffs for buyers and sellers in each round as a percentage of the maximum total payoffs, which we explain in section 4.4.

Table 5 reports the participant-level tests next to the session-level exact tests for the four outcomes that are defined for a participant: seller delivery and buyer entry, by buyer type. The estimates are the same in both analyses, and only the p-value differs. They disagree on one contrast, type a buyer entry ($p = 0.030$ with participants as units versus 0.200 with sessions). type a delivery does not differ ($p = 0.769$ and 0.933), and neither does any type b contrast.

4.2 Trustworthiness

Type a trustworthiness is 15.4% under Verified Review and 14.7% under Fake Review (Table 4 and Figure 2; $p = 0.933$). The four Verified Review sessions range from 0% to 44%, and the two Fake Review sessions are at 19% and 10% (Table A.1. Fake reviews therefore did not significantly lower type a trustworthiness relative to verified reviews, which does not support the prediction (H2) that trustworthiness is lowest under Fake Review in type a. type

[3]The sessions ran in date order by treatment: No Review on April 4,5, and 6, Verified Review on April 7, 8, 9, and 12, and Fake Review on April 13, and 14, 2022. Treatment is therefore confounded with the date of the session. Any difference between the groups of sessions that has nothing to do with the review system, such as who signed up or how the lab ran those days, would look like a treatment effect, and our tests cannot detect it. All nine sessions took place within nine weekdays, which makes a seasonal or semester-level change unlikely, but it does not remove the concern. The two Fake Review sessions were also the last two run and the two largest, with 22 and 18 subjects. Our comparisons between treatments should be read with this in mind.

[4]As a check on dependence within sessions, we ran Kruskal–Wallis tests, which rank the participant means (after removing differences between treatments) and ask whether participants differ across the nine sessions; we find no such difference for any of the four outcomes. The p-values are 0.466 for type a trust, 0.762 for type b trust, 0.256 for type a trustworthiness and 0.845 for type b trustworthiness.

Table 4 Fake versus Verified Review: exact permutation tests on session means

Outcome		Verified Review	Fake Review	Fake − Verified	Exact two-sided p
Trust (%)	Type A	45.9	26.3	-19.7	0.200
		(12.1)	(15.0)		
	Type B	65.8	70.9	5.1	0.733
		(10.5)	(1.2)		
	A − B	-19.9	-44.6	-24.7	0.133
		(11.2)	(13.8)		
Trustworthiness (%)	Type A	15.4	14.7	-0.7	0.933
		(19.7)	(6.6)		
	Type B	86.6	76.4	-10.2	0.133
		(5.4)	(3.1)		
	A − B	-71.2	-61.7	9.5	0.667
		(24.4)	(9.7)		
Efficiency (%)	Type A	30.5	48.7	18.2	0.133
		(8.3)	(12.9)		
	Type B	76.5	76.2	-0.3	0.933
		(4.0)	(0.4)		
	A − B	-46.0	-27.5	18.5	0.067
		(6.8)	(12.5)		

The unit of observation is the session (4 Verified Review, 2 Fake Review). The test statistic is the difference in mean session-level outcomes; the p-value enumerates all 15 reassignments of these six sessions to the two treatments, so it is at least $1/15 = 0.067$. No adjustment for multiple testing is made. Standard deviations across sessions are in parentheses below the means (four Verified Review sessions, two Fake Review sessions). A − B is the within-session difference between Type A and Type B buyers. Trustworthiness is conditional on the buyer entering. Comparisons with No Review are in Appendix 3.

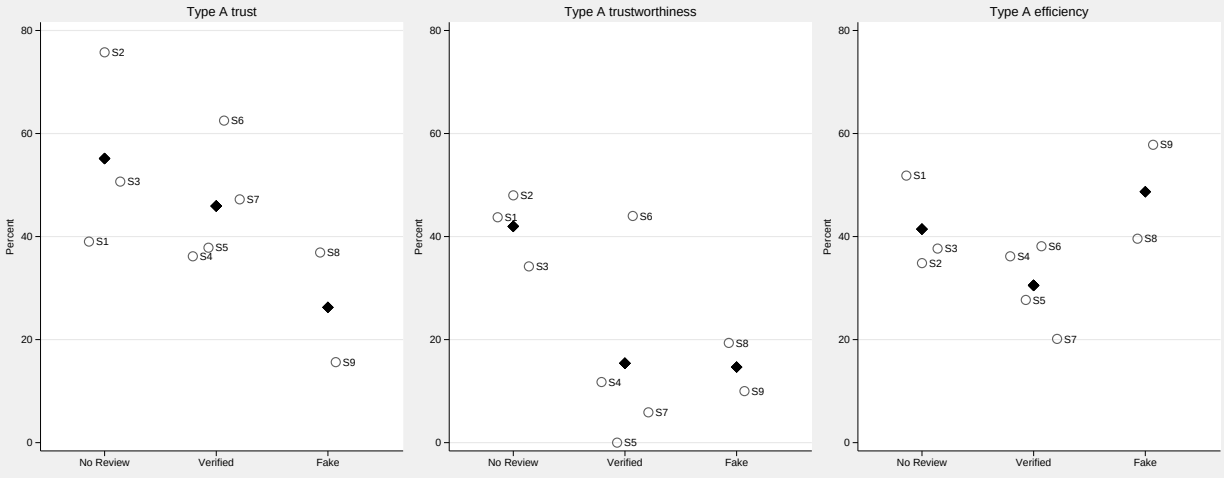


Figure 2 Session means for type a buyers

Note: Each circle is one session, labeled as in Table A.1. Diamonds are the means of the session means in each treatment.

Table 5 Fake versus Verified Review: participant-level Mann–Whitney tests

Outcome		Verified Review	Fake Review	Fake − Verified	p	N (Verified/Fake)
Trust (%)	Type A	48.0	28.3	-19.7	0.030	20/20
		(27.8)	(25.0)	(8.4)		
	Type B	63.9	67.9	4.0	0.828	20/20
		(29.2)	(24.7)	(8.6)		
Trustworthiness (%)	Type A	17.8	16.3	-1.5	0.769	20/17
		(29.4)	(30.0)	(9.8)		
	Type B	86.1	77.3	-8.8	0.428	19/20
		(13.8)	(25.7)	(6.6)		

The unit of observation is the participant: each participant contributes the mean of his or her own rounds (seller delivery is defined only over rounds in which a buyer of that type entered). The treatment columns give the mean of these participant means, in percentage points, with the standard deviation across participants in parentheses below each mean; because each participant counts once, these means differ slightly from the means of session means in Table 4, where each session counts once. The standard error of the difference, in parentheses below it, is $\sqrt{s_F^2/n_F + s_V^2/n_V}$, treating participants as independent.

b trustworthiness is high under both (86.6% and 76.4%), and the lower Fake Review value is not statistically distinguishable from Verified Review ($p = 0.133$), which supports the prediction on type b trustworthiness. Fake reviews had little room to lower trustworthiness, because verified reviews give sellers little reason to deliver (discussed in Section 5.2).

Result 1*: Fake reviews did not lower type a trustworthiness relative to verified reviews (14.7% versus 15.4%), and verified reviews themselves delivered little discipline. type b trustworthiness is not significantly different in the two review sessions.*

Review versus no review. H1 predicts that verified reviews raise type a trustworthiness above the no-review baseline. The data show the opposite: type a trustworthiness is 42.0% without reviews and 15.4% under Verified Review. The difference is significant with participants as units (Mann–Whitney $p = 0.031$, Table A.7) but not with sessions as units (exact $p = 0.114$, Table A.5), so H1 is not supported for type a buyers. type b trustworthiness is 87.0% without reviews and 86.6% under Verified Review ($p = 0.914$ with sessions as units), as H1 predicts. The type a gap is not robust across specifications: in the last six rounds, delivery is 22.6% without reviews and 16.7% under Verified Review, and with sessions as units the pooled p-value for the two review treatments ranges from 0.036 to 0.476 across nine variants (Table A.8).

Result 2*: Verified reviews did not raise type a trustworthiness relative to No Review: it is lower, significantly so with participants as units but not with sessions as units, and the gap is not robust across specifications. The trustworthiness of Type b remains unchanged.*

4.3 Trust

In all treatments, trust has a downward trend in early rounds and levels off in later rounds (Figure A.1). type a participation is 45.9% in Verified Review and 26.3% in Fake Review (Table 4). Neither this contrast nor type b participation (65.8% and 70.9%, $p = 0.733$) is statistically significant at the session level: $p = 0.200$ for type a participation. Type a contrast has $p = 0.030$ with participants as units (Table 5). The direction is that of H3, lower trust under Fake Review for type a and no difference for type b.

Most of the difference comes from one session. Type a participation is 15.6% in session S9 but 36.9% in S8, which is within the range of the four Verified Review sessions (36.2–62.5%), and the two Fake Review sessions do not differ significantly from each other (Mann–Whitney $p = 0.158$). This lower participation explains why type a efficiency is higher in the Fake Review sessions: the Fake Review treatment "protects" type a buyers from untrustworthy sellers.

The ratings displayed in the Fake Review sessions were inflated by seller-posted ratings, which are higher than the buyers' own feedback (Table 7), so the exit cannot be attributed to visibly low ratings. Opting out is the payoff-improving choice in every treatment. A type a buyer who enters expects $18\gamma - p_a$ against an outside option of 3, and delivery is far below the 0.94 that makes entry pay in every session. Type a buyers who entered lose about 11 points in both review treatments (Table A.12, Panel D, which also gives the No Review values).

The puzzle is therefore why so many buyers enter, not why fewer do under Fake Review. Four patterns bear on the extra exit:

1. Buyers who stayed out tend to stay out, and more so in Fake Review: 17.8% enter

after staying out under Fake Review, against 27.4% under Verified Review (Table A.12 Panel C).

2. Exit is also present in the first paid rounds, before ratings can have accumulated: 50.0% of type a buyers enter in paid rounds 1 and 2 under Fake Review, against 75.0% under Verified Review (Table A.12 Panel A; cells of 18 and 20 buyer-rounds).
3. The extra exit is not explained by being cheated. Fewer type a buyers in the Fake Review sessions were cheated at least once in rounds 1 to 8, although buyers who enter less often have fewer chances to be cheated, and the share who enter in a round after being cheated in their previous type a round is similar in the two treatments (Table A.12 Panel C).
4. Buyers discount high ratings in Fake Review: At a displayed rating of 4 or above, 33.3% enter under Fake Review against 80.0% under Verified Review, although only ten Verified Review observations have such a rating (Table A.12 Panel B).

The second and third patterns fit buyers who distrust ratings that they know can be faked.

Result 3*: Type a trust is lower in Fake Review than Verified Review sessions (26.3% against 45.9%), which is significant at participant-level but not session level. Type b trust is slightly lower in Fake Review but is not significant in both tests.*

4.4 Market Efficiency and Payoffs

Market efficiency is calculated by maximum payoff divided by total payoffs. For type a buyer, the maximum payoff is 8, which equals the maximum value gained by buyers, 18, minus the seller's cost of providing a good a, 10. Similarly, the maximum total payoff is 14 in type b buyer cases, which equals to 18 minus the cost of providing a good b, 4.

Type a market efficiency is 30.5% in the Verified Review sessions and 48.7% in the Fake Review sessions, and type b efficiency is 76.5% and 76.2% (Table 4). The results do not show that fake reviews lower market efficiency; notably, the Fake Review treatment exhibits higher type a efficiency than the Verified Review treatment, although the difference is not statistically significant across sessions ($p = 0.133$), and the within-session type a minus type b efficiency gap has $p = 0.067$, the smallest value attainable with these sessions. As the theory in Section 3.1 suggests, the total payoff depends solely on trustworthiness and trust. Lower trustworthiness leads to a lower total payoff and, thus, lower market efficiency. The relationship between trust and market efficiency specifically depends on whether the market contains a sufficient proportion of trustworthy sellers.

Result sections 4.2 and 4.3 explain that the efficiency gain from the Fake Review treatment for type a therefore seems to come from buyers choosing *Out*.

Result 4*: Fake Review sessions show higher type a efficiency than Verified Review sessions (48.7% versus 30.5%), but the difference is not statistically significant between sessions (exact $p = 0.133$), and it reflects exit by type a buyers.*

In terms of buyer and seller payoffs, type a buyers lose money in both treatments (-3.4 points per round under Verified Review and -0.8 under Fake Review, -10.8 points under Verified Review and -11.1 under Fake Review when choosing to enter), while type b buyers earn 5.7 and 7.1; sellers earn 5.8 and 4.7 from type a buyers and 5.0 and 3.5 from

type b buyers (Table A.12). Seller's payoff is higher in the verified review treatment and lower in the fake review treatment, while buyer payoff has the opposite trend. This suggests that sellers have more bargaining power in verified review sessions, while buyers do better in fake review sessions. However, no contrast is statistically significant at the session level (the smallest p-value is 0.133).

Since type b efficiency is the same in the two treatments (76.5% and 76.2%), the higher payoff for buyers and the lower payoff for sellers in type b is a transfer between them, which the lower price that type b buyers pay under Fake Review accounts for.

A posted price is the price a seller sets for type a or type b buyers in a round, before the buyer's type is known. Paid prices are the realized prices in transactions in which the buyer entered, and they affect the distribution of total payoffs between buyers and sellers (Appendix Table A.6). Sellers posted higher prices in the verified review than in the fake review treatment, especially for type b. Paid prices are 14.49 and 14.76 for type a and 10.90 and 9.17 for type b. None of these differences are statistically significant. A seller's price does not predict delivery to type a buyers (-1.1 points per price point, standard error 0.8; Table A.10).

Result 5*: Buyer and seller payoffs do not differ significantly between Fake and Verified Review. Buyers do better and sellers worse under Fake Review in type b, and entering type a buyers lose about 11 points in both treatments.*

Result 6*: Sellers post and receive somewhat lower prices from type b buyers under Fake Review than under Verified Review, but no price difference is statistically significant.*

4.5 Reviews

Reviews can matter for the above outcomes only if buyers use them and sellers respond to them. Table 6 summarizes two findings on buyers and sellers' response to average ratings.

1. Buyers respond to ratings. Among type a buyers, each additional point in the displayed rating (the average of past ratings, on the 1–5 scale) increases the probability of entering by 9.1 percentage points under Verified Review (standard error 5.1) and 7.1 under Fake Review (standard error 2.0; Table 6). Type b buyers do not respond. Buyers therefore use ratings where it matters. Section 5.2 asks whether this gives sellers a good enough reason to deliver.

2. Sellers do not manage their ratings. Within the seller, delivery to type a buyers does not respond significantly to the displayed rating: the coefficients are 6.0 and 10.1 percentage points per rating point with standard errors of 9.1 and 27.1 (Table 6). We find no evidence that sellers deliver more when their rating is low or milk it when it is high. Sellers also deliver less to type a buyers than to type b buyers in every session (Table A.1), but this also appears without reviews, where the cost motive alone predicts it. The type-contingent pattern is, therefore, not evidence of strategic rating management.

Result 7*: Buyers respond to the displayed rating, but sellers' delivery does not respond to their displayed rating.*

Table 6 Response to the displayed rating (percentage points per rating point)

	Trust (buyer enters)		Trustworthiness (seller delivers)	
	Verified Review	Fake Review	Verified Review	Fake Review
Type A buyers	9.1* (5.1)	7.1*** (2.0)	6.0 (9.1)	10.1 (27.1)
Type B buyers	5.1 (6.0)	-0.8 (2.8)	3.3 (9.5)	0.1 (7.1)
Observations (Type A/Type B)	148/147	140/160	63/98	37/114

Each coefficient is from a separate linear regression on the rating displayed to the buyer for the matched seller (average of past ratings, 1–5). Trust regressions use buyer rows and include session and round fixed effects; trustworthiness regressions use seller rows in rounds in which a buyer of that type entered and include seller and round fixed effects. Standard errors in parentheses are clustered by participant and understate dependence within sessions. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

4.6 Fake Reviews in the Fake Review Sessions

Tables 7 and A.4 provide a breakdown of the occurrence of fake reviews, by type and by round, indicating whether they were present ("Yes") or absent ("No"), along with the average rating associated with each category. The prevalence of fake reviews among sellers reveals a notable reliance on such practices to influence both their own ratings and those of their peers. The sellers in the two Fake Review sessions used fake reviews extensively, with 84.4% of the seller-rounds including a self-assigned rating and 72.2% of the seller-rounds rating for another seller. As anticipated, sellers tend to assign higher ratings to themselves than to others, averaging 4.70 out of 5 for self-reviews and 4.02 for reviews of others. Only about half of the seller-rounds have real feedback from buyers, which is less often than fake reviews, and both types of fake ratings are higher than the real feedback. Appendix 2 describes fake ratings in two regressions in Fake Review sessions (Tables **??** and **??**). In the first, sellers who posted a higher fake self-rating in the previous round delivered less often: each rating point goes with a 14-percentage-point lower probability of delivering ($p < 0.01$, 107 seller-rounds, with and without seller fixed effects), while fake ratings for another seller show no relation. The second asks who posts fake ratings: sellers with a low displayed rating are more likely to post a fake self-rating when the buyer is type a ($p < 0.05$), and the other coefficients are at most marginally significant. The rating displayed in these sessions mainly reflects how much the sellers post (Table A.11).

A seller's mean displayed rating is strongly related to the share of rounds in which the seller posted a fake self-rating (Spearman correlation 0.72 across 20 sellers) than to the feedback buyers gave (0.12), and sellers who posted fake self-ratings more often delivered less often to type a buyers (Spearman correlation -0.49, $p = 0.052$ across 17 sellers). This is the pattern in Mayzlin (2006), where firms spend more resources promoting inferior products. As a result, the displayed rating points the wrong way across sellers: delivery to entered type a buyers is 6.9% (2 of 29) when the displayed rating was 4 or above and 37.5% (3 of 8) below 4, and the negative relation between the displayed rating and delivery is no longer distinguishable from zero once the seller's posting share is controlled for (Table A.11), while

type a buyers' entry rises with the displayed rating. Buyers who follow the rating are directed toward sellers who post more and deliver less. In Verified Review, there is no such pattern (Table A.11).

Table 7 Incidence of Fake Review

Review Type	Yes	No	Avg. Rating (Yes)
Fake Reviews			
Self	270 (84.4%)	50 (15.6%)	4.70
Other	231 (72.2%)	89 (27.8%)	4.02
Real Feedback			
Real Feedback (Fake Review)	163 (50.9%)	157 (49.1%)	3.29
Real Feedback (Verified Review)	181 (56.6%)	139 (43.4%)	2.70

Sample: sellers only. Fake reviews are sellers provided, while real feedback is from buyers.

Result 8*: Sellers in the Fake Review sessions posted fake ratings in most rounds. The displayed rating mostly reflects how much a seller posts, and sellers who posted more fake self-ratings delivered less often to type a buyer.*

5 Discussion

5.1 Conclusion

In this paper, we examined the economic impact of fake reviews in a laboratory market for experience goods, focusing on market efficiency, trustworthiness, trust, and the payoff distribution between buyers and sellers. The evidence supports one conclusion, offers an explanation for it, and leaves one question open.

The conclusion is that in our environment, fake reviews did not reduce trustworthiness relative to verified reviews, but did decrease trust. Trustworthiness was low under both review systems, so fake reviews had little room to lower it, and their effect appears in what buyers do, not in what sellers do. Verified reviews also guide buyers better than fake reviews: across sellers, a higher displayed rating goes with lower delivery in the Fake Review sessions, while under Verified Review there is no such pattern (Table A.11).

5.2 Mechanism

The reputational incentive constraint says when reviews would discipline sellers: a seller delivers the costly good to a type a buyer only if the reputational gain is at least the 6-point cost. We ask whether this gain is large enough, using the within-session evidence on how buyers respond to ratings in Section 4.5.

One delivery barely moves the rating. To see why, follow one seller who faces an entering type a buyer under Verified Review, step by step (Table 8).

1. **The cost.** Delivering the high-quality good costs the seller 6 points more than cheating.
2. **The rating after the transaction.** The buyer's rating depends only a little on what the seller did. Type a buyers who received the high-quality good rated the seller 3.07 on average, and type a buyers who were shortchanged rated 1.71. Delivering raises the rating by 1.36 points.
3. **The displayed average.** Buyers see the average of all ratings the seller has received so far, which is about 4.8 ratings. One more rating that is 1.36 points higher raises that average by about $1.36/(4.8 + 1) = 0.23$ points.
4. **The buyers' response.** Each extra point in the displayed average raises a type a buyer's probability of entering by 9.1 percentage points (Table 6). A 0.23-point rise therefore raises it by about 2.1 percentage points.
5. **The value of the extra entries.** The seller expects to meet about 4.2 more type a buyers (48% of the 8.6 rounds that remain on average), and each additional entry is worth about 7.3 points (the price, minus the cost of the low-quality good, minus the seller's outside option of 3 points). The expected gain is about 0.64 points.
6. **The comparison.** The seller gives up 6 points to receive about 0.64 points back, so delivering loses about 5.4 points in expectation.

Under Fake Review each step is weaker. The rating after delivery averages 2.29, against 1.74 after cheating, a gap of 0.55. Seller-posted ratings lengthen the history to about 16 ratings, so one delivery changes the average by 0.03 points, which raises type a entry by 0.2 percentage points. The expected gain is about 0.07 points. The calibration is an upper bound, because it assumes the change in the rating persists undiluted over the remaining rounds.

Even if ratings were perfectly informative, with a rating of 5 after every delivery and 1 after every case of cheating, the gain would rise only to 1.9 points under Verified Review and 0.5 under Fake Review.

Sellers do not buy ratings either. Sellers also supply the costly high-quality good to buyers who do not need it, and they do so more often in the Fake Review sessions. Type b buyers are satisfied by the cheap good, so supplying the costly one wastes 6 points. It happens in 14.8% of entered type b transactions under Verified Review and 23.8% under Fake Review (Table 9; 13.4% without reviews). One possible motivation is that sellers are buying ratings. The data do not support it. Over-delivery earns almost nothing: a type b buyer rated a seller who over-delivered 3.31 against 3.16 under Verified Review and 3.90 against 3.74 under Fake Review, differences of 0.22 and 0.14 points (standard errors 0.32 and 0.18), for a cost of 6 points.

In the Fake Review sessions sellers supply the high-quality good more often to type b buyers (23.8%) than to type a buyers (17.1%). This is the opposite of what a cost motive or generosity would predict. It looks like heuristic or noisy behavior, not a rating strategy,

Table 8 Reputational return to delivering the high-quality good to a Type A buyer

	Verified Review	Fake Review
Buyer's rating after delivery (1–5)	3.07	2.29
Buyer's rating after cheating (1–5)	1.71	1.74
Number of ratings behind the displayed average	4.8	16.1
Change in the displayed rating from one delivery (points)	0.23	0.03
Buyers' response (percentage points of Type A entry per rating point)	9.1	7.1
Gain in type a entry from one delivery (percentage points)	2.1	0.2
Number of type a buyers the seller can still expect to meet	4.2	4.4
Value to the seller of one additional entry (points)	7.3	7.2
Reputational gain from delivering (points)	0.64	0.07
Cost of delivering (points)	6	6

Rough, upper-bound calibration of the reputational incentive constraint using the review sessions. The change in the displayed rating is the feedback difference divided by one plus the number of ratings behind the average. Buyers' response is the Type A coefficient in Table 6. The gain assumes that the change in the rating persists undiluted over the remaining rounds. The value of an additional entry is the average posted Type A price in the entered Type A transactions (14.27 under Verified Review and 14.17 under Fake Review) minus the cost of the low-quality good (4) minus the seller's outside option (3): the seller is assumed to give the additional buyer the cheap good and to receive the outside option of 3 points if that buyer would have stayed out. The gain from delivering is the gain in Type A entry times the number of Type A buyers still to come times the value of an additional entry.

Table 9 Supplying the high-quality good to buyers who do not need it

	No Review	Verified Review	Fake Review
Over-delivery to Type B buyers (% of entered transactions)	13.4	14.8	23.8
Delivery to Type A buyers (% of entered transactions)	40.5	19.2	17.1
Type B buyer's rating of the seller after over-delivery			
After over-delivery	–	3.31	3.90
After appropriate delivery	–	3.16	3.74
Difference (standard error)	–	0.22 (0.32)	0.14 (0.18)
Over-delivery per one-point higher posted p_b(percentage points)	-1.1 (0.7)	-0.5 (0.8)	-1.5 (0.8)

Over-delivery means the seller supplied the costly high-quality good to a Type B buyer, who is satisfied by the cheap good; it costs the seller 6 points. Shares are pooled over entered transactions. The rating comparison uses session fixed effects, and the price regressions use session and round fixed effects; standard errors in parentheses are clustered by seller. Sellers within a session are not independent, so counts of sellers are descriptive.

although a seller could hold the mistaken belief that over-delivery buys ratings, and this experiment cannot tell.

In our environment, reviews did not discipline sellers because the reputational return to honesty is small. Buyers respond to ratings, but one delivery moves a seller's rating by 0.23 points under Verified Review and 0.03 under Fake Review, and the implied gain is roughly 11% and 1% of the cost of delivering. This is the logic of Aperjis and Johari (2010): when a score averages a long window and many transactions are cheap to satisfy, the score stays high whatever the seller does in the costly cases. In the terms of the reputational incentive constraint, reviews fail because the condition fails by roughly an order of magnitude, and fake reviews make this worse, because seller-posted ratings lengthen the history that dilutes each delivery and carry no information about delivery.

5.3 Related Experiments

Our results sit between two strands of the experimental literature. First, the weak response to ratings resembles G. E. Bolton et al. (2019), where sellers react less to noisy feedback and ship lower quality. Our ratings are noisy in the same way: a type a buyer rates a seller only 1.4 points higher after delivery than after cheating.

Second, previous reputation experiments generally find that feedback helps. None finds lower trustworthiness with a review system than without one (G. Bolton et al., 2004; Solimine & Isaac, 2023). Our No Review comparison (Appendix 2) points the other way, although it is not statistically robust. The closest study, Krügel and Paetzel (2024), finds that costless rating fraud lowers the reliability of displayed ratings, expectations, contributions, and efficiency relative to a control in which fraud is impossible. We find insignificant difference in type a trustworthiness between Fake and Verified Review, and higher type a efficiency because type a buyers exit. A likely reason is the reputational incentive constraint of our game is not satisfied: honesty is costly only with type a buyers, and the buyer decides whether to enter, so fake ratings show up as exit, not as lower delivery.

Theory also warns against reading our dilution result too broadly. The length of the rating history need not limit efficiency (Dellarocas, 2005), its effect on incentives runs in both directions depending on the seller (Aperjis & Johari, 2010), and averaging can raise cooperation (Greiff & Paetzel, 2020). The feedback in G. Bolton et al. (2004) is over 30 rounds, and buyers in Solimine and Isaac (2023) choose among competing sellers. Our setting differs on these points: ratings are noisy buyer scores, buyers can only accept or reject the matched seller, and the game lasts 16 rounds.

5.4 Implications and Limits

The open question is whether fake reviews change participation and efficiency relative to verified reviews. The Fake Review sessions show lower type a participation and higher efficiency with the participant-level evidence. Because type a entry lowers total surplus when trustworthiness is below 5/6, which holds in every session, the higher efficiency reflects exit, not better outcomes. The efficiency gain is buyers protecting themselves, not the review system working as intended. We cannot say whether buyers exit because they distrust ratings that they know can be faked. Buyers who stay follow the ratings, although in these sessions

the sellers who posted more fake self-ratings delivered less (Table A.11).

For platforms, this means a review system, even a verified one, is not enough on its own: ratings must also carry enough weight and distinguish delivery from cheating well enough, to make honesty pay. Catching fake reviews is only part of the problem.

The result is specific to our environment. Subjects knew that there were exactly 16 paid rounds, matching was random within a session, a seller's rating was a single average that dilutes each delivery, and delivering cost 6 points against about 7 points for an additional entry. The experiment cannot say whether fake reviews would lower trustworthiness in a market where reputation does discipline sellers, for example, with a longer horizon, a rating that weights recent transactions more heavily, or a lower cost of honesty. The results should not be read as showing that review systems fail to discipline sellers in general.

The study has limits. The sessions were run in date order by treatment, No Review on April 4 to 6, Verified Review on April 7 to 12, and Fake Review on April 13 and 14, 2022, so treatment is confounded with the date of the session, and the two Fake Review sessions were also the last two runs and the two largest, with 22 and 18 subjects (Section 4). Ratings are a single number rather than the mix of stars and text buyers see on real platforms, and buyers are re-matched with a new seller every round. Sixteen paid rounds is short compared to how long a reputation takes to build or unwind on a real platform.

References

Akerlof, G. A. (1978). The market for "lemons": Quality uncertainty and the market mechanism. In *Uncertainty in economics* (pp. 235–251). Elsevier.

Aperjis, C., & Johari, R. (2010). Optimal windows for aggregating ratings in electronic marketplaces. *Management Science*, *56*(5), 864–880. https://doi.org/10.1287/mnsc.1090.1145

Aronow, P. M., Baron, J., & Pinson, L. (2019). A note on dropping experimental subjects who fail a manipulation check. *Political Analysis*, *27*(4), 572–589.

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

Bolton, G., Greiner, B., & Ockenfels, A. (2013). Engineering trust: Reciprocity in the production of reputation information. *Management science*, *59*(2), 265–285.

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

Bolton, G. E., Kusterer, D. J., & Mans, J. (2019). Inflated reputations: Uncertainty, leniency, and moral wiggle room in trader feedback systems. *Management Science*, *65*(11), 5371–5391. https://doi.org/10.1287/mnsc.2018.3191

Bowles, S., & Polanía-Reyes, S. (2012). Economic incentives and social preferences: Substitutes or complements? *Journal of Economic Literature*, *50*(2), 368–425. https://doi.org/10.1257/jel.50.2.368

Chen, D. L., Schonger, M., & Wickens, C. (2016). Otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, *9*, 88–97.

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

Darby, M. R., & Karni, E. (1973). Free competition and the optimal amount of fraud. *The Journal of law and economics*, *16*(1), 67–88.

Dellarocas, C. (2005). Reputation mechanism design in online trading environments with pure moral hazard. *Information Systems Research*, *16*(2), 209–230. https://doi.org/10.1287/isre.1050.0054

Dellarocas, C. (2006). Strategic manipulation of internet opinion forums: Implications for consumers and firms. *Management Science*, *52*(10), 1577–1593. https://doi.org/10.1287/mnsc.1060.0567

Dellarocas, C., & Wood, C. A. (2008). The sound of silence in online feedback: Estimating trading risks in the presence of reporting bias. *Management science*, *54*(3), 460–476.

Dimoka, A., Hong, Y., & Pavlou, P. A. (2012). On product uncertainty in online markets: Theory and evidence1. *MIS quarterly*, *36*(2), 395–426.

Ert, E., Fleischer, A., & Magen, N. (2016). Trust and reputation in the sharing economy: The role of personal photos in airbnb. *Tourism management*, *55*, 62–73.

Frey, B. S., & Oberholzer-Gee, F. (1997). The cost of price incentives: An empirical analysis of motivation crowding-out. *American Economic Review*, *87*(4), 746–755.

Gazzale, R. S., & Khopkar, T. (2011). Remain silent and ye shall suffer: Seller exploitation of reticent buyers in an experimental reputation system. *Experimental Economics*, *14*(2), 273–285. https://doi.org/10.1007/s10683-010-9267-z

Greiff, M., & Paetzel, F. (2020). Information about average evaluations spurs cooperation: An experiment on noisy reputation systems. *Journal of Economic Behavior & Organization*, *180*, 334–356.

He, S., Hollenbeck, B., & Proserpio, D. (2022). The market for fake reviews. *Marketing Science*.

Huck, S., Lünser, G. K., & 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.

Keser, C., & Späth, M. (2021). The value of bad ratings: An experiment on the impact of distortions in reputation systems. *Journal of Behavioral and Experimental Economics*, *95*, 101782.

Klein, B., & Leffler, K. B. (1981). The role of market forces in assuring contractual performance. *Journal of Political Economy*, *89*(4), 615–641.

Klein, L. R. (1998). Evaluating the potential of interactive media through a new lens: Search versus experience goods. *Journal of business research*, *41*(3), 195–203.

Kreps, D. M., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*(2), 245–252.

Krügel, J. P., & Paetzel, F. (2024). The impact of fraud on reputation systems. *Games and Economic Behavior*, *144*, 329–354. https://doi.org/10.1016/j.geb.2024.01.013

Lal, R., & Sarvary, M. (1999). When and how is the internet likely to decrease price competition? *Marketing Science*, *18*(4), 485–503.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, *76*(3), 1071–1102.

Louis, P., Markopoulos, P., Voucharas, G., & Xefteris, D. (2025). *Market performance and review system informativeness* [Working paper, University of Cyprus. Presented at CRETE 2025]. https://pages.aueb.gr/conferences/Crete2025/papers/Louis.pdf

Luca, M. (2011). Reviews, reputation, and revenue: The case of yelp. com (no. 12-016). *Harvard Business School*.

Luca, M., & Zervas, G. (2016). Fake it till you make it: Reputation, competition, and yelp review fraud. *Management Science*, *62*(12), 3412–3427.

Lumeau, M., Masclet, D., & Pénard, T. (2010). Les conséquences de la manipulation de la réputation dans les systèmes d'évaluation en ligne. une étude expérimentale. *Revue économique*, *61*(6), 1123–1133. https://doi.org/10.3917/reco.616.1123

Lynch, M., Miller, R. M., Plott, C. R., & Porter, R. (1991). Product quality, informational efficiency, and regulations in experimental markets. *Research in experimental economics*, *4*, 269–318.

Mavlanova, T., Benbunan-Fich, R., & Koufaris, M. (2012). Signaling theory and information asymmetry in online commerce. *Information & management*, *49*(5), 240–247.

Mayzlin, D. (2006). Promotional chat on the internet. *Marketing Science*, *25*(2), 155–163.

Mayzlin, D., Dover, Y., & Chevalier, J. (2014). Promotional reviews: An empirical investigation of online review manipulation. *American Economic Review*, *104*(8), 2421–55.

Montgomery, J. M., Nyhan, B., & Torres, M. (2018). How conditioning on posttreatment variables can ruin your experiment and what to do about it. *American Journal of Political Science*, *62*(3), 760–775. https://doi.org/10.1111/ajps.12357

Mudambi, S. M., & Schuff, D. (2010). What makes a helpful online review? a study of customer reviews on amazon. com. *MIS quarterly*, *34*(1), 185–200.

Mukherjee, A., Venkataraman, V., Liu, B., & Glance, N. (2013). What yelp fake review filter might be doing? *Proceedings of the international AAAI conference on web and social media*, *7*(1), 409–418.

Nelson, P. (1970). Information and consumer behavior. *Journal of political economy*, *78*(2), 311–329.

Nelson, P. (1974). Advertising as information. *Journal of political economy*, *82*(4), 729–754.

Peck, J., & Childers, T. L. (2003). To have and to hold: The influence of haptic information on product judgments. *Journal of marketing*, *67*(2), 35–48.

Plott, C. R., & Wilde, L. L. (1980). *Professional diagnosis vs. self-diagnosis: An experimental examination of some special features of markets with uncertainty*. California Institute of Technology, Division of the Humanities; Social Sciences.

Rayana, S., & Akoglu, L. (2015). Collective opinion spam detection: Bridging review networks and metadata. *Proceedings of the 21th acm sigkdd international conference on knowledge discovery and data mining*, 985–994.

Resnick, P., Kuwabara, K., Zeckhauser, R., & Friedman, E. (2000). Reputation systems. *Communications of the ACM*, *43*(12), 45–48.

Shapiro, C. (1983). Premiums for high quality products as returns to reputations. *Quarterly Journal of Economics*, *98*(4), 659–679.

Solimine, P., & Isaac, R. M. (2023). Reputation and market structure in experimental platforms. *Journal of Economic Behavior & Organization*, *205*, 528–559.

Appendix Supplementary Tables and Figures

Appendix 1 Supplementary Figures

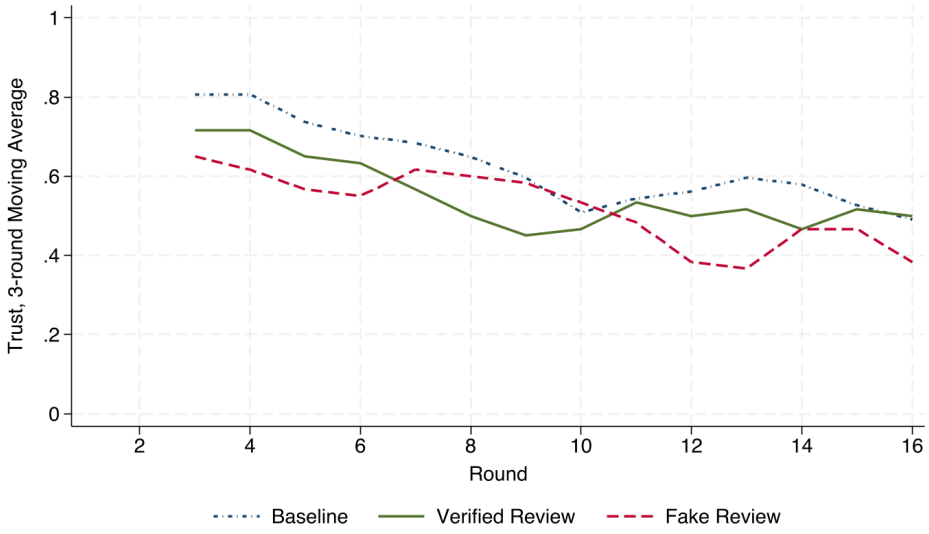


Figure A.1 3-round Moving Average of Trust from Buyers, Baseline vs. Verified Review vs. Fake Review Treatments

Appendix 2 Supplementary Tables

Table A.1 Session-level outcomes by buyer type

			Trust (%)		Trustworthiness (%)		Efficiency (%)	
Treatment	Session	Subjects	Type A	Type B	Type A	Type B	Type A	Type B
No Review	S1	10	39.0	64.1	43.8	88.0	51.8	76.2
	S2	10	75.8	72.3	48.0	88.2	34.8	80.5
	S3	18	50.7	76.8	34.2	84.9	37.7	81.8
Mean of sessions			55.1	71.1	42.0	87.0	41.4	79.5
Verified Review	S4	10	36.2	72.7	11.8	83.3	36.2	79.2
	S5	8	37.8	51.9	0.0	92.9	27.7	70.9
	S6	12	62.5	75.0	44.0	81.0	38.1	79.6
	S7	10	47.2	63.6	5.9	89.3	20.1	76.3
Mean of sessions			45.9	65.8	15.4	86.6	30.5	76.5
Fake Review	S8	22	36.9	71.7	19.4	74.2	39.6	75.9
	S9	18	15.6	70.0	10.0	78.6	57.8	76.4
Mean of sessions			26.3	70.9	14.7	76.4	48.7	76.2

Each cell is the mean of one session's 16 paid rounds (Trust: buyer enters; Trustworthiness: seller delivers the type-appropriate good, conditional on entry; Efficiency: pair payoff as a percentage of the maximum). Session labels are anonymous. "Mean of sessions" weights each session equally.

Table A.2 Summary statistics of participants

	No Review			Verified Review			Fake Review		
	N	Mean	SD	N	Mean	SD	N	Mean	SD
Female	38	0.68	0.47	40	0.68	0.47	40	0.57	0.50
Age	37	20.22	2.39	40	20.45	1.77	40	20.68	1.76
Asian	38	0.26	0.45	40	0.17	0.38	40	0.33	0.47
Black	38	0.53	0.51	40	0.62	0.49	40	0.57	0.50
Hispanic	38	0.03	0.16	40	0.03	0.16	40	0.05	0.22
Multiracial	38	0.05	0.23	40	0.03	0.16	40	0.03	0.16
White	38	0.11	0.31	40	0.12	0.33	40	0.03	0.16
Prefer not to answer	38	0.03	0.16	40	0.03	0.16	40	0.00	0.00

One observation per participant (118 participants: 38, 40, and 40 in the three treatments). Female and the race categories are 0/1 indicators; age is in years, and one participant in the No Review sessions has no valid age.

Table A.3 Fake versus Verified Review: regressions with one observation per participant and round

Outcome	Verified Review mean	Fake Review coefficient	Standard error	p	Observations (participants)
Buyer: Type A entry (%)	45.6	-19.1	8.8	0.036	308 (40)
Buyer: Type B entry (%)	67.5	4.5	6.9	0.517	332 (40)
Seller: Type A delivery (%)	19.2	-3.6	10.4	0.735	114 (37)
Seller: Type B delivery (%)	85.2	-9.9	6.9	0.161	230 (39)

Linear probability models in percentage points, estimated on paid rounds of the four Verified Review and two Fake Review sessions, with round fixed effects. Entry uses buyer rows of the given type; delivery uses seller rows in rounds in which a buyer of that type entered. The coefficient is the difference between Fake and Verified Review. Standard errors are clustered by participant. Treatment varies only across sessions, so these standard errors understate dependence within sessions and the p-values rest on the same independence assumption as the participant-level tests; session-clustered standard errors are not reported because there are only six sessions, two of them Fake Review. The session-level exact tests in Table 4 are the primary inference.

Table A.4 Fake rating incidence by round, sellers in the Fake Review sessions

	Self fake rating		Other fake rating		
Round	No	Yes	No	Yes	Total
1	4	16	4	16	20
2	3	17	3	17	20
3	2	18	5	15	20
4	5	15	8	12	20
5	2	18	5	15	20
6	2	18	5	15	20
7	5	15	7	13	20
8	3	17	7	13	20
9	5	15	7	13	20
10	2	18	6	14	20
11	3	17	7	13	20
12	1	19	4	16	20
13	2	18	6	14	20
14	4	16	5	15	20
15	4	16	7	13	20
16	3	17	3	17	20
Total	50	270	89	231	320

Each round has the 20 sellers of the two Fake Review sessions; buyers cannot post fake ratings. Self: a seller posted a rating for itself; Other: a seller posted a rating for another seller.

Table A.5 Exact permutation tests on session means, all contrasts

		Mean of session means			Exact two-sided p-value			
Outcome		No Review	Verified	Fake	VR vs. Base	FR vs. VR	FR vs. Base	Review vs. Base†
Efficiency (%)	Type A	41.4	30.5	48.7	0.257	0.133	0.500	0.607
	Type B	79.5	76.5	76.2	0.371	0.933	0.300	0.167
	A − B	-38.1	-46.0	-27.5	0.514	0.067	0.500	0.905
Trust (%)	Type A	55.1	45.9	26.3	0.400	0.200	0.200	0.214
	Type B	71.1	65.8	70.9	0.486	0.733	1.000	0.560
	A − B	-15.9	-19.9	-44.6	0.743	0.133	0.100	0.345
Trustworthiness (%)	Type A	42.0	15.4	14.7	0.114	0.933	0.100	0.060
	Type B	87.0	86.6	76.4	0.914	0.133	0.100	0.381
	A − B	-45.1	-71.2	-61.7	0.171	0.667	0.100	0.107
Buyer payoff	Type A	-2.1	-3.4	-0.8	0.314	0.133	0.400	0.762
	Type B	6.3	5.7	7.1	0.314	0.200	0.300	0.869
Seller payoff	Type A	5.4	5.8	4.7	0.571	0.133	0.300	1.000
	Type B	4.8	5.0	3.5	0.714	0.133	0.100	0.607

The unit of observation is the session (3 No Review, 4 Verified Review, 2 Fake Review). The test statistic is the difference in mean session-level outcomes; p-values enumerate every reassignment of sessions to treatments (35, 15, 10 and 84 assignments), so each p-value is at least one over the number of assignments. No adjustment for multiple testing is made. A − B is the within-session difference between Type A and Type B buyers. Trustworthiness is conditional on the buyer entering. †Pools Verified and Fake Review sessions; chosen after seeing the data.

Table A.6 Posted and paid prices: exact permutation tests on session means

	Mean of session means			Exact two-sided p-value			
Price	No Review	Verified	Fake	VR vs. Base	FR vs. VR	FR vs. Base	Review vs. Base†
Posted price, Type A	14.19	14.77	14.42	0.143	0.200	0.500	0.119
Posted price, Type B	10.90	11.51	10.21	0.371	0.267	0.200	0.774
Paid price, Type A	13.94	14.49	14.76	0.514	0.733	0.500	0.429
Paid price, Type B	10.40	10.90	9.17	0.486	0.133	0.100	0.940

Prices are in points. A posted price is the price a seller sets for Type A or Type B buyers in a round; a paid price is the posted price in transactions in which the buyer entered. The unit of observation is the session; the tests are as in Table A.5. †Pools Verified and Fake Review sessions; chosen after seeing the data.

Table A.7 Participant-level tests, all contrasts

		Mean of participant means (SD)			*p*-value (Mann–Whitney)				Participants
Outcome		No Review	Verified	Fake	VR vs. Base	FR vs. VR	FR vs. Base	Review vs. Base†	(NR/VR/FR)
Trust (%)	Type A	52.5	48.0	28.3	0.517	0.030	0.017	0.081	19/20/20
		(30.9)	(27.8)	(25.0)					
	Type B	71.3	63.9	67.9	0.662	0.828	0.778	0.929	19/20/20
		(15.6)	(29.2)	(24.7)					
Trustworthiness (%)	Type A	36.4	17.8	16.3	0.031	0.769	0.022	0.008	19/20/17
		(31.1)	(29.4)	(30.0)					
	Type B	87.9	86.1	77.3	0.269	0.428	0.141	0.143	19/19/20
		(21.2)	(13.8)	(25.7)					

The unit of observation is the participant: each participant contributes the mean of his or her own rounds (seller delivery is defined only over rounds in which a buyer of that type entered), and the tests are Mann–Whitney tests. The treatment columns give the mean of these participant means, in percentage points, with the standard deviation across participants in parentheses below each mean. The tests assume that no shock is shared by all sessions of a treatment.

Table A.8 Robustness of the session-level tests

Variant	Sessions (B/V/F)	No Review	Verified	Fake	p: Review vs. Base	p: Fake vs. Verified
Panel A. Delivery of the costly good to Type A buyers (%)						
1. All sessions and rounds (as in the paper)	3/4/2	42.0	15.4	14.7	0.060	0.933
2. Drop the last three rounds	3/4/2	43.8	16.6	14.6	0.060	0.933
3. Drop the first session run	2/4/2	41.1	15.4	14.7	0.071	0.933
4. Rounds 7–16 only	3/4/2	34.7	14.2	11.8	0.107	1.000
5. Drop sellers who over-delivered in round 1	3/4/2	42.7	20.6	7.7	0.131	0.800
6. Drop payoff-dominated decisions	3/4/2	41.6	15.0	13.6	0.036	0.933
7. Drop all sellers who ever over-delivered	3/4/2	45.4	21.7	3.3	0.119	0.533
8. Last six rounds only (11–16)	3/4/2	22.6	16.7	7.1	0.440	0.600
9. Drop all sellers who ever posted a price below 4	3/4/2	31.9	20.4	9.1	0.476	0.933
Panel B. Targeting index (costly share to Type A minus to Type B, points)						
1. All sessions and rounds (as in the paper)	3/4/2	29.0	2.0	-8.9	0.024	0.467
2. Drop the last three rounds	3/4/2	28.7	0.1	-9.2	0.024	0.533
3. Drop the first session run	2/4/2	27.7	2.0	-8.9	0.071	0.467
4. Rounds 7–16 only	3/4/2	25.0	8.0	-2.5	0.119	0.467
5. Drop sellers who over-delivered in round 1	3/4/2	31.9	7.4	-10.0	0.060	0.400
6. Drop payoff-dominated decisions	3/4/2	27.1	1.9	-13.2	0.024	0.400
7. Drop all sellers who ever over-delivered	3/4/2	45.4	21.7	3.3	0.119	0.533
8. Last six rounds only (11–16)	3/4/2	9.5	15.4	-15.1	0.786	0.133
9. Drop all sellers who ever posted a price below 4	3/4/2	12.5	8.2	-0.7	0.798	0.867
Panel C. Type A buyer entry (%)						
1. All sessions and rounds (as in the paper)	3/4/2	55.1	45.9	26.3	0.214	0.200
2. Drop the last three rounds	3/4/2	58.2	48.0	30.4	0.190	0.200
3. Drop the first session run	2/4/2	63.2	45.9	26.3	0.179	0.200
4. Rounds 7–16 only	3/4/2	48.5	36.8	20.2	0.214	0.267
5. Drop sellers who over-delivered in round 1	3/4/2	54.9	44.5	25.3	0.179	0.200
6. Drop payoff-dominated decisions	3/4/2	57.5	47.0	26.1	0.202	0.133
7. Drop all sellers who ever over-delivered	3/4/2	52.8	41.3	28.4	0.214	0.400
8. Last six rounds only (11–16)	3/4/2	50.7	34.5	13.7	0.155	0.200
9. Drop all sellers who ever posted a price below 4	3/4/2	58.8	46.7	30.4	0.250	0.267
Panel D. Efficiency gap, Type A minus Type B (points)						
1. All sessions and rounds (as in the paper)	3/4/2	-38.1	-46.0	-27.5	0.905	0.067
2. Drop the last three rounds	3/4/2	-38.9	-46.6	-31.8	0.798	0.200
3. Drop the first session run	2/4/2	-44.9	-46.0	-27.5	0.464	0.067
4. Rounds 7–16 only	3/4/2	-36.6	-37.6	-22.5	0.667	0.200
5. Drop sellers who over-delivered in round 1	3/4/2	-37.9	-40.9	-29.6	0.952	0.333
6. Drop payoff-dominated decisions	3/4/2	-34.4	-46.2	-24.0	0.631	0.067
7. Drop all sellers who ever over-delivered	3/4/2	-33.1	-41.2	-38.5	0.345	0.667
8. Last six rounds only (11–16)	3/4/2	-41.5	-38.9	-13.1	0.429	0.133
9. Drop all sellers who ever posted a price below 4	3/4/2	-51.4	-49.5	-39.4	0.821	0.600

Each row repeats the session-level analysis after the stated change; p-values are two-sided exact permutation tests. Variants 5, 7 and 9 drop sellers, and variant 6 drops decisions, by their own choices, and a dropped seller or decision removes the whole pair-round, so they select on behavior and are descriptive (Table A.9 bounds them). Variant 6 drops pair-rounds in which the seller posted a price below 4 (a loss even with the cheap good) or a Type B buyer entered at a price above 15 (a certain loss against the outside option).

Table A.9 Trimming bounds for delivery to Type A buyers under selection

Comparison	Retained share (first arm / second arm)	Trimmed share	Retained-only difference	Lee bounds (points)	95% interval for the bounds
Panel A. Retained = never gave the costly good to a Type B buyer					
Review (pooled) vs. No Review	0.42 / 0.68	0.38	-25.2	[-45.5, -2.6]	[-73.3, 12.3]
Verified vs. No Review	0.45 / 0.68	0.34	-20.2	[-37.5, 0.8]	[-71.6, 22.0]
Fake vs. No Review	0.40 / 0.68	0.42	-30.8	[-53.2, -6.7]	[-83.8, 6.7]
Fake vs. Verified	0.40 / 0.45	0.11	-10.6	[-12.3, -4.8]	[-40.8, 6.7]
Panel B. Retained = never posted a price below 4					
Review (pooled) vs. No Review	0.53 / 0.32	0.40	-14.2	[-30.4, -3.5]	[-59.4, 47.5]
Verified vs. No Review	0.55 / 0.32	0.43	-9.5	[-30.4, 6.0]	[-59.4, 64.8]
Fake vs. No Review	0.50 / 0.32	0.37	-19.4	[-30.4, -13.0]	[-59.4, 30.3]
Fake vs. Verified	0.50 / 0.55	0.09	-9.9	[-12.0, -2.0]	[-50.0, 21.2]
Panel C. Retained = the Type A buyer entered (selection into entry; unit = match)					
Review (pooled) vs. No Review	0.37 / 0.53	0.30	-22.1	[-39.6, 3.6]	[-76.6, 24.0]
Verified vs. No Review	0.46 / 0.53	0.14	-21.3	[-27.9, -11.7]	[-58.8, 15.2]
Fake vs. No Review	0.28 / 0.53	0.48	-23.4	[-60.5, 17.1]	[-91.4, 28.6]
Fake vs. Verified	0.28 / 0.46	0.39	-2.1	[-14.5, 17.1]	[-48.3, 29.3]

In Panels A and B the outcome is each seller's rate of delivering the costly good to Type A buyers, and a seller is retained if the exclusion rule does not apply and the outcome is observed. In Panel C the unit is a match between a seller and a Type A buyer, the outcome (delivery) is observed only if the buyer entered, and a match is retained if the buyer entered; the bounds assume that reviews only reduce entry, and the bootstrap resamples sellers. Because the rule removes different shares of sellers in different arms, the arm that retains more sellers is trimmed from the top or bottom of its outcome distribution by the difference in retention, which gives bounds on the effect for sellers who would be retained in both arms (Lee, 2009). Intervals are percentile intervals from 4,000 bootstrap draws that resample sellers within arms. Dropping participants by post-treatment behavior can bias treatment comparisons (Aronow et al., 2019; Montgomery et al., 2018).

Table A.10 Delivery to entered Type A buyers: the review gap with price and subject controls

	(1)	(2)	(3)
Review sessions (vs. No Review), points	-23.5**	-23.5**	-27.5**
	(11.0)	(11.0)	(10.3)
Price posted for Type A, points per price point		-1.1	
		(0.8)	
Gender indicator (1 vs. 0), points			6.7
			(8.2)
Age, points per year			5.8
			(4.1)
Round fixed effects	Yes	Yes	Yes
Entered Type A transactions	193	193	193
Sellers	56	56	56

Linear probability models of delivering the costly good to an entered Type A buyer, in percentage points. Standard errors in parentheses are clustered by seller and understate dependence within sessions. Chosen after seeing the data. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.11 The displayed rating, fake ratings, delivery and entry for Type A buyers

	Verified Review	Fake Review
Delivery to entered Type A buyers, by the rating displayed to the buyer		
Rating 4 or above: delivered / entered	1/8 (12.5%)	2/29 (6.9%)
Rating below 4: delivered / entered	11/55 (20.0%)	3/8 (37.5%)
Difference (points)	-7.5	-30.6
Fisher exact p	1.000	0.057
Percentage points per rating point		
Seller delivery, across sellers (session and round fixed effects)	-4.0	-46.7***
	(4.8)	(6.7)
Seller delivery, within seller (seller and round fixed effects)	6.0	10.1
	(9.1)	(27.1)
Buyer entry (session and round fixed effects)	9.1*	7.1***
	(5.1)	(2.0)
Fake Review sessions: sellers who post fake self-ratings (seller level)		
Spearman correlation, share of rounds with a fake self-rating and delivery rate to entered Type A buyers (17 sellers)	–	-0.49 (p =0.052)
Above-median posters: delivered / entered	–	3/29 (10.3%)
Below-median posters: delivered / entered	–	4/12 (33.3%)
Fisher exact p, transactions as units	–	0.165
Spearman correlation, mean displayed rating and posting share (20 sellers)	–	0.72
Spearman correlation, mean displayed rating and mean buyer feedback received	–	0.12
Delivery per displayed rating point across sellers, controlling for posting share	–	-17.8 (19.5)

The displayed rating is the average of past ratings on the 1–5 scale, shown to the buyer once the seller has a rating, from the second round of each phase on. In the Fake Review sessions it includes ratings that sellers posted for themselves and for other sellers. The cutoff of 4 was chosen after seeing the data. Linear probability models in percentage points per rating point; standard errors in parentheses are clustered by participant and understate dependence within sessions. The Fisher test treats transactions as independent and is descriptive. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

Table A.12 Type A buyers: entry, experience and payoffs

	No Review	Verified Review	Fake Review
Panel A. Entry by round (% of Type A buyer-rounds, count)			
Paid rounds 1–2	75.0% (9/12)	75.0% (15/20)	50.0% (9/18)
Paid rounds 3–6	64.3% (27/42)	57.6% (19/33)	33.3% (11/33)
Paid rounds 7–16	45.3% (43/95)	36.4% (39/107)	21.6% (21/97)
Panel B. Entry by the rating displayed to the buyer			
Rating 4 or above	–	80.0% (8/10)	33.3% (29/87)
Rating below 4	–	39.9% (55/138)	15.1% (8/53)
Panel C. Entry after the buyer's previous Type A round (%, count)			
Previous round: cheated	57.1% (24/42)	50.0% (27/54)	48.4% (15/31)
Previous round: received the costly good	76.7% (23/30)	84.6% (11/13)	28.6% (2/7)
Previous round: stayed out	31.0% (18/58)	27.4% (20/73)	17.8% (16/90)
Cheated at least once in rounds 1–8 (% of buyers)	68% (n=19)	90% (n=20)	60% (n=20)
Panel D. Payoffs (points); the outside option is 3			
Payoff of a Type A buyer who entered	-6.8	-10.8	-11.1
Mean payoff per Type A buyer-round	-2.2	-3.3	-0.9

Type A buyers, paid rounds. Panel B uses the review sessions, where a rating is displayed once the seller has one. Panel C conditions on the buyer's previous round as a Type A buyer. The cells are small and the sessions are not independent, so the numbers are descriptive and were chosen after seeing the data.