Good bargains and reputable sellers

An experimental investigation of electronic feedback systems

Domenico Colucci (University of Florence - Italy)
Simone Salotti (National University of Ireland, Galway)
Vincenzo Valori (University of Florence - Italy)

This version: March 2011

Abstract

Electronic reputation mechanisms aim at signaling the quality of traders and at hampering misbehavior. When a buyer rates a seller, the main aspect to keep into account should be the consistency between the good's advertised characteristics and the actual good itself. On the other hand, the buyer's satisfaction may also be related to the surplus stemming from the transaction, i.e. whether or not the purchase was a good bargain. This constitutes a possible source of distortion in the way sellers are rated. Using an experimental investigation, we find that the transaction payoff is in fact a significant factor driving the feedback. Conversely, the match (or mismatch) between the ex-ante advertised description and the item received is comparatively less important, and may be offset by the effect of the payoff. This result accounts for the occurrence on some e-marketplace contexts of sellers getting away with inflated descriptions of the goods on sale without compromising their reputation.

Keywords: Reputation systems, feedback behavior, electronic markets, electronic commerce

1. Introduction

Electronic commerce relationships differ from those established in standard, face-to-face, markets. In electronic markets there is a higher uncertainty, since most transactions occur among agents that have never met and repeated interactions are rare. Both the identity of electronic trading parties and the product's quality are uncertain (Ba and Pavlou 2002), standard legal channels to solve disputes are hard to implement (Bolton et al. 2004), and there are more problems with fraud (Gregg and Scott 2006). To mitigate risks, electronic reputation mechanisms based on feedback have been established, attracting the interest of many scholars in the last decade. The reference literature for this kind of studies goes back to Akerlof (1970), since information asymmetries can lead to opportunism and even market failures. Feedback systems try to circumvent these problems, helping to build trust among the potential trading parties (Jarvenpaa et al. 2000, Moon and Kim 2001, Gefen et al. 2003, Kohen 2003, Pavlou 2003, Josang et al. 2006). Such systems however have not yet been able to eliminate internet auction fraud (see Gavish and Tucci 2008).

Most of the literature on feedback systems focuses on the effects of sellers' feedback scores either on the probability of selling (Yang et al. 2007) or on the price premium obtained (Melnik and Alm, 2002, Lucking-Reiley et al. 2007, Houser and Wooders 2006, Standifird 2001), or both (McDonalds and Slawson 2002, Resnick and Zeckhauser 2002, Dellarocas 2003, Livingston 2005, Zhang 2006). It is now generally agreed that feedback profiles affect both prices and the probability of sale, but the evidence on the precise effects is mixed. However, it seems that the adverse effects of receiving a negative feedback are higher than the favorable effects of a positive feedback. Moreover, while early studies mainly deal with quantitative measures of positive and negative feedback (such as the total number of positive ratings, gross or net of the negative ones, or the percentage of positive ratings over the total), some of the most recent ones focus on the comments that complement the feedback (Utz et al. 2009). There is some previous experimental work. Bolton et al. (2004) and Gazzale and Khopkar (2011) use a two stage game to study the effects of reputation on market efficiency, while another strand of literature (starting with Keser 2003) uses trust games experiments mainly to investigate the strategic components of reporting behavior.

While the perspective in the literature has chiefly been that of measuring the consequences of online feedback, the present paper aims at reversing the point of view, trying to understand its causes. In turn this relates to whether feedback systems can actually achieve their purposes of countering informational asymmetries and favoring truthful descriptions of goods on sale, thereby hampering misbehavior and frauds. Although feedback determinants are obviously unobservable, the literature on customer satisfaction offers, in our view, a useful guidance.[1] Previous research (Anderson and Sullivan 1993; Inman et al. 1997; Rust et al. 1999) suggests the importance of disconfirmation ("disappointment" in Inman et al. 1997) in determining the post-transaction customer satisfaction. Disconfirmation is the difference between expected and realized outcomes. In electronic commerce, expectations are largely driven by the information provided by the sellers. Therefore, if disconfirmation is the sole determinant of feedback, then the reputation system can

[1] The relationship among satisfaction, word of mouth and feedback behavior has been studied by e.g. Anderson (1998) and Söderlund (1998).

effectively fulfill its purpose: for example, a significant mismatch between the actual delivered good and what the customer expected, results, through the lever of disconfirmation, in the negative rating of an untrustworthy seller.

However, can we rule out other factors as causes of the feedback decisions, such as whether the transaction was a good bargain? In principle, one could expect rating to be unrelated to the transaction surplus obtained by the buyer, in particular when the seller does not control the price, as in an auction. But is it actually so?

Whether or not buyers are able to produce "unbiased" ratings constitutes the main focus of this paper. We use a laboratory investigation to inquire into the determinants of positive and negative feedback given by the buyers. Such factors are difficult to observe using field evidence and they are hard to elicit precisely even in interviews with online shoppers or auction bidders. On the contrary, an experimental context allows to control for the true value which accrues to the buyers and to make ex-ante beliefs on sellers as unbiased as possible. In the present work this comes at the cost of various limitations: on one hand we concentrate only on the demand side, while on the other we solicit the buyers' responses to an oversimplified supply side. Due to the urge to control variables such as the prices, the design entails a very basic trading context[3] and avoids the complications of auction mechanisms to allocate the goods.

Our results indicate that disconfirmation is not the unique determinant of feedback behavior. In fact, the transaction surplus, proxied within the experiment by the economic payoff, appears to be even more important. Put differently, awarding a good bargain in our experiment seems to strongly compensate for disconfirmation, thus leading to biased feedback profiles. Besides, a form of "positive rating bias" arises, by which benevolent rating of bad sellers is more common than negative rating of good ones.

The rest of this paper is organized as follows. Section 2 motivates the analysis and explains the rationale of the experiment. Section 3 contains the details and the objectives of the experimental design. Section 4 lays out results. Section 5 concludes.

2. Motivation and theory

The principal aim of our investigation is to make a step forward in identifying the wedge between what online feedback on sellers should be about from what it actually is about. Assuming normative principles are complied with, the reputation mechanism should be able to produce reliable feedback profiles, stimulating trust and trustworthiness among the traders. However, is such socially optimal behavior borne out by the evidence?

We assume that feedback ratings are to a certain extent linked to the satisfaction experienced by the buyers. Thus, we take advantage of the literature on satisfaction determinants as discussed in the Introduction. In particular, we hypothesize that the difference between the expectations and the realized outcome of the transaction (disconfirmation - see Rust et al. 1999; Anderson and Sullivan 1993, Inman et al. 1997) is an

[3] In such trading context there is no scope for other empirically relevant factors such as politeness and efficiency of communication, timeliness and packaging.

important determinant of the feedback behavior. We also conjecture that both the transaction surplus and the existing feedback profile of the seller play a non-negligible role.[4] Such factors, if found to have significant momentum, would introduce a bias in the way sellers are rated. Formally, we posit the following feedback function:

$$fb_{ijt} = f(disconfirmation, surplus,\ fb_{jt-1}), \qquad (1)$$

where fb_{ijt} is the feedback (positive/negative) left by buyer i to seller j at time t; $fb_{jt\text{-}1}$ is the past cumulated feedback of the seller. To see how this model can generate specific research questions within our experimental design consider its building blocks as follows.

Our experimental subjects, all buyers, have to buy fictional items sold by a pool of virtual sellers and contribute to build a reputation system about the sellers based on a feedback mechanism. Subjects are aware that sellers are computer-generated and that half of them are "good" and the other half are "bad", in the sense that buyers will on average be better off trading with the former than with the latter. The experiment is built around a between-subjects design consisting of three treatments. The key element in which they differ lies in the different signals that encode the type of seller, indicating whether it is a good or a bad seller. The signals are linked to disconfirmation in treatment 1, to the transaction payoff in treatment 2, and to a third, artificial factor in treatment 3. The subjects do not know the type of seller for sure before purchasing, as the signals are disclosed only after the transaction has taken place. Then, the subjects are asked to leave a feedback (positive or negative). For the buyers it would be ideal to know the sellers' type ex-ante: such knowledge can in principle be replaced effectively by the reputation mechanism, once it is up and running.

If subjects coordinate on rating good sellers positively and bad sellers negatively (which we label henceforth as socially optimal behavior) then they will be able to anticipate the seller's type perfectly once a seller has sold at least once. More specifically, this means that, when complying with socially optimal behavior, they should:

- in treatment 1, rate according to disconfirmation (ignoring the payoff);
- in treatment 2, rate according to payoff (ignoring the disconfirmation);
- in treatment 3, rate according to whether the good's value is an odd or even number (ignoring both disconfirmation and payoff).

The different treatments allow us to understand to what extent the subjects are able to use the signals as prescribed by the above socially optimal behavior. Conversely, we can check the capacity to resist disconfirmation and/or payoff when they act as confounding stimuli. Also, we aim at assessing the relative importance of our variables of interest in determining the feedback behavior.

We investigate the experimental data along two dimensions: i) the efficiency of the feedback mechanism and the comparison with the socially optimal behavior; ii) the identification of the determinants of feedback. Overall the analysis allows addressing the following questions: to what extent do the explanatory variables of equation (1) play a role in determining feedback in the experimental setting? Will the resulting feedback

[4] That the exposure to prior ratings may influence the rating behavior of later buyers has been put forward as early as in Asch (1956).

profiles permit to distinguish between the types of sellers? Are there significant treatment effects in terms of the relative efficiency of the feedback mechanism? Is the observed behavior significantly different from the socially optimal benchmark? If so, how are the “wrong” ratings (negative to good sellers and positive to bad ones) distributed among the two types of sellers?

3. Experimental design

3.1 Basic structure and treatments

The structure of the experiment is as follows. All subjects play the role of buyers and perform two tasks at each round. First, they have to buy a fictional item in a market, then they must leave a feedback about the seller, thus contributing to build a reputation system. Sellers are computer-generated; half of them are “good” (named “type A”) and the other half are “bad” (named “type B”) sellers, in a probabilistic sense which is explained precisely below. There is no strategic interaction among the subjects and leaving a negative feedback after a transaction should not be grounded in punishment/anger motives, but in the need to ease the identification of a bad seller in future transactions.[5] Subjects are enabled to distinguish type A from type B sellers once the transaction has been carried out. The initial profiles of the sellers are blank, so the feedback system emerges entirely endogenously from the experimental subjects’ behavior.

At the beginning of each round, every subject is matched with three potential sellers, each offering an item whose value is advertised to be equal to 100 experimental points.[6] The experiment is built so as to guarantee that such advertised value can be considered a correct unconditional expectation for the value of the item (whereas conditioning on the type of seller, which is ex-ante unknown, returns different expectations). Each seller is characterized by information regarding its feedback profile and the price it proposes (which is randomly generated for each round and seller).

A buyer must choose from which one of the three sellers to buy. Then, the true value of the item bought is revealed and the type of seller can be unambiguously identified. In general the true value differs from the advertised value. Their difference, which we label as the “lie”, is our experimental counterpart of disconfirmation. The payoff obtained in the transaction is equal to the difference between the true value of the item and the price paid. The subjects cannot incur in losses: a “satisfied or reimbursed” clause is at work, so that if the true value of the item is lower than the price paid, the seller “refunds” the buyer, and the payoff from the transaction is simply zero. The rationale for this is to avoid loss aversion effects (Novemsky and Kahneman 2005), as well as bankruptcy issues or biased incentives due to higher initial show up fees. Therefore, we decided to leave the investigation of losses to further research. Finally, the buyer must rate the

[5] In most online marketplaces only sellers are rated, therefore there is no scope for strategic feedback behavior. This holds even on eBay, where feedback is bilateral but sellers cannot rate buyers negatively. This rule was introduced in 2007 to avoid retaliatory behavior: previous to that, subjects who misbehaved could avoid negative feedback by threatening to leave negative feedback in return. See Klein et al. (2009) for an analysis of the consequences of this change.

[6] All relevant variables are denominated in experimental points, which are then converted to Euros according to a rate which is treatment specific.

seller by leaving a feedback of the type "positive/negative" which is incorporated in the profile of the seller starting from the following round.

The experimental design is between-subjects and it involves three treatments whose defining factor consists in the signals that unequivocally identify the seller's type. Socially optimal behavior calls for exploiting the signals to identify the sellers' type and rating them consistently leaving a positive feedback for types A and a negative feedback for types B.[7] In the *first treatment* type A sellers deliver items whose true value is a random integer drawn from the interval between 101 and 120. On the contrary, type B sellers will cheat buyers, delivering items whose true value is an integer between 80 and 100 (see Figure 1 for details). Thus, the difference between the true value and the advertised value of the item, the lie (which can in fact be hostile or pleasant), unambiguously determines the type of seller. Because it is possible to earn little money from a type A seller, or much from a type B, the payoff may act as a confounding stimulus with respect to the socially optimal behavior.

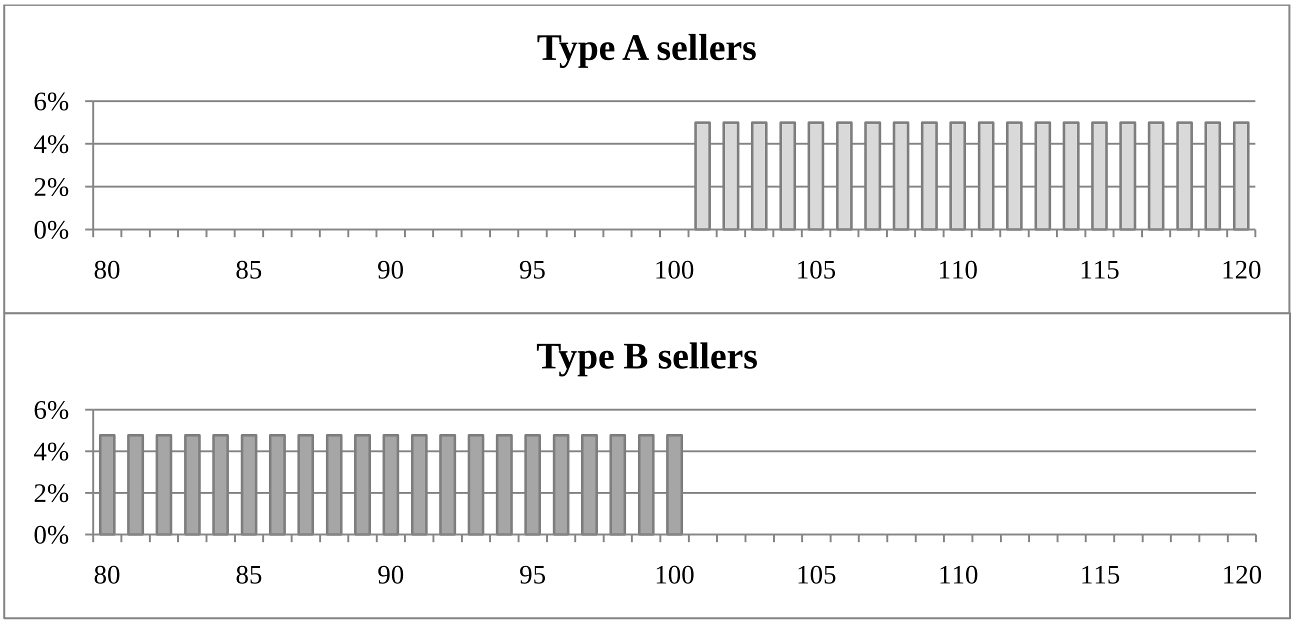

Figure 1: probability distribution of true values, treatment 1

In the *second treatment* type A sellers deliver items whose true value is an integer between 21 and 40 points higher than the price paid, independent of the advertised value. Type B sellers deliver items whose true value is an integer between 0 and 20 points in excess of the price paid, again independently from the advertised value (see Figure 2 for details). Thus, the difference between the true value of the item and the price paid (the payoff from the transaction) is the unambiguous signal to identify the seller type. Note that this is independent of the lie told by the seller, since a type A seller can lie as well as a type B seller can provide a

[7] In principle the rating system is a way of signaling which could also be used inverted: but in this case given the obvious semantic connotation of positive/negative feedbacks we can safely stick to the more reasonable scheme in which types A are rated positively, and types B negatively.

good worth more than the advertised value. Thus, buyers should ignore the advertised value altogether. The lie is a confounding variable in this case.

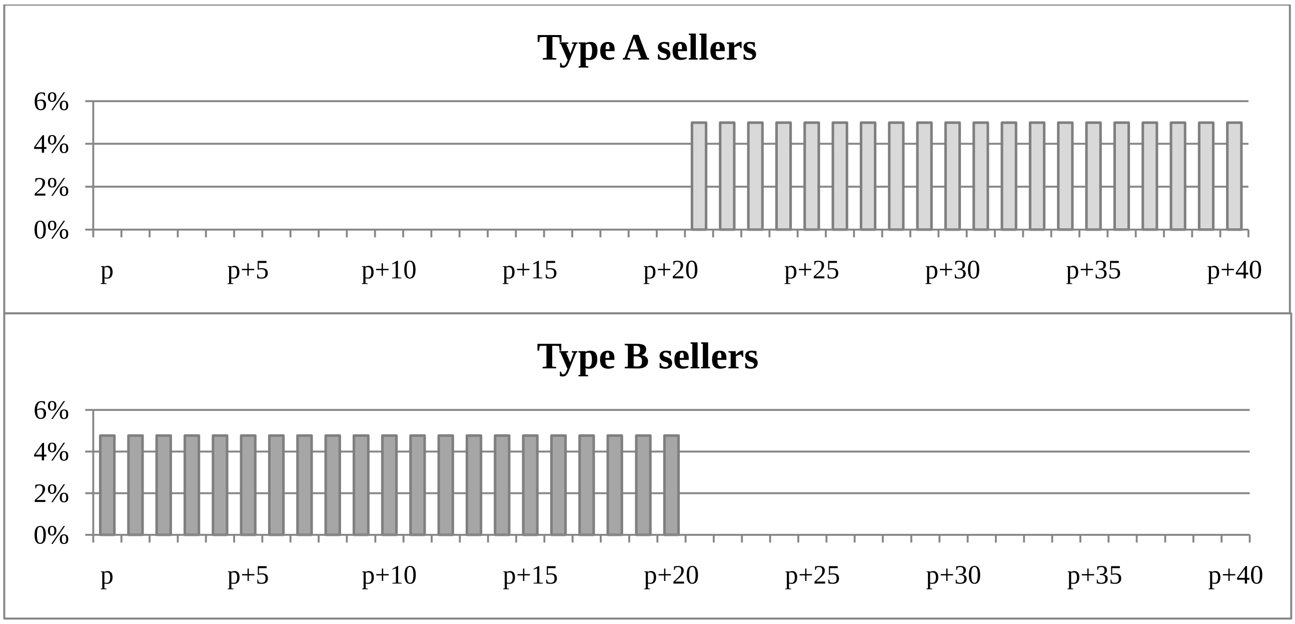

Figure 2: probability distribution of true values, treatment 2

In the *third treatment* type A sellers deliver high true value items more often than type B sellers and such true values are always even numbers, while type B sellers always deliver odd true values items. Figure 3 shows the probability distribution for the true values delivered by the sellers. Basically, type A sellers provide better-than-advertised items with a higher probability with respect to the types B (75% vs. 25%), who in turn will rip-off buyers more often than types A.

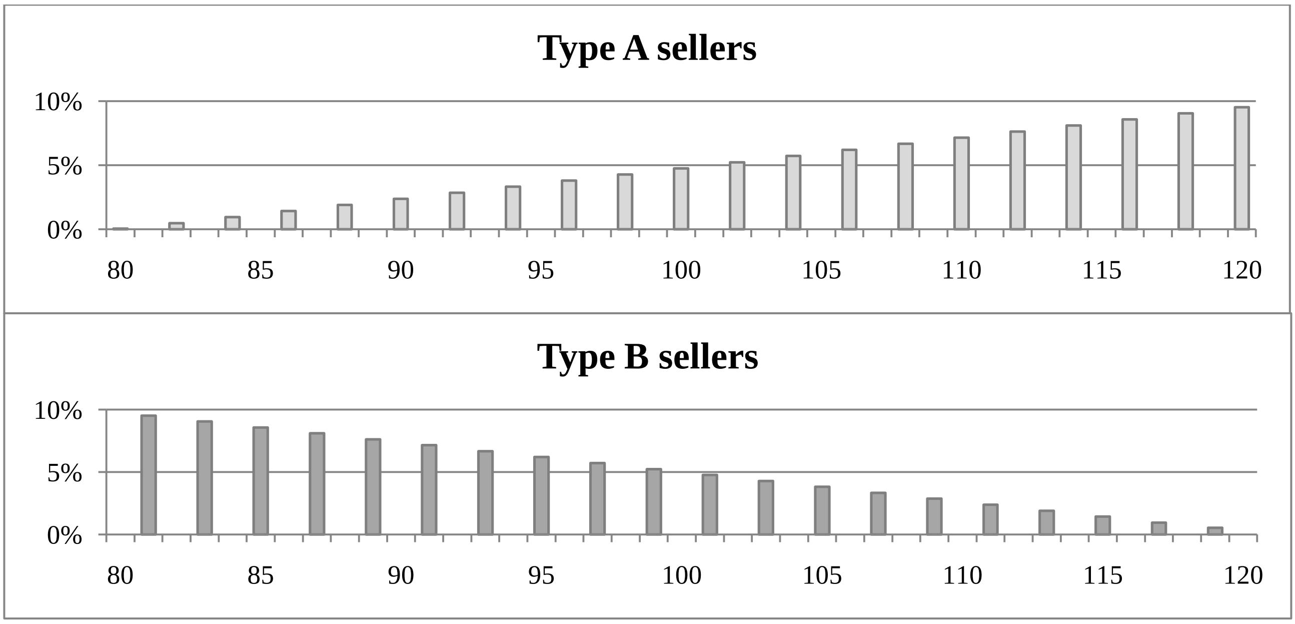

Figure 3: probability distribution of true values, treatment 3

In this case the socially optimal behavior requires to check whether true value is even or odd and leave the feedback consistently, no matter the observed payoff or the sign/size of the lie involved in the transaction. Note that in this case type A sellers are better (both in terms of payoffs provided and lies told) than types B probabilistically (e.g., the expected payoff when buying from the former is 26, as opposed to 13 from the latter), but not in each and every case. Both lie and payoff are confounding stimuli in this case given that type A (type B) sellers can provide scarce (high) payoffs and/or lie hostilely (pleasantly). Table 1 summarizes how the signals to identify the seller's type differ across the three different treatments.

Table 1. Seller's type signals in the different treatments

	Treatment 1	Treatment 2	Treatment 3
Lie	Yes	No	No
Payoff	No	Yes	No
Even/Odd	No	No	Yes

It is interesting to see under which of the treatments the feedback system is more efficient (in terms of minimizing the distance between the socially optimal profiles and those generated through the experiment).

3.2 The experimental protocol

Each treatment takes place over 30 rounds. At each round, each buyer has to buy a fictional item whose advertised value is 100 throughout the whole length of the experiment. At the beginning of each round, 30 buyers are matched with three sellers selected from a pool of 45 computer-generated sellers (22 types A and 23 types B). The choice is probabilistic, so that the probability of appearing in the triples presented to each buyer is inversely related to the number of transactions carried out in the past (that is, the more a seller has been chosen by the buyers in the past, the lower its probability to appear again among the possible choices). The reason for this is that we want to collect a significant number of observations about both types. As a by-product, the advertised value of the item corresponds to the ex ante expected true value that buyers can receive from the set of sellers. This is ensured by this transaction-based mechanism of sellers' selection. The subjects were only told that the selection of the sellers was random.

Each seller is characterized by the following information:

- name (an alpha-numeric string of 4 characters);
- history/feedback profile (number of transactions effected, number of positive and negative ratings received, positive feedback rate);
- price (variable across rounds and sellers).

Note that, in each round, the same seller can be associated with zero, one or more buyers. In the latter case, the price at which it sells is not necessarily the same across subjects (the same goes for the true value of the item). The price is a randomly generated integer in the set of multiples of 5 between 60 and 100, using the discrete uniform distribution. Once the buyer has chosen from which seller to buy the item, she discovers its true value (with the constraint that in case of price exceeding the true value the payoff is set to zero). Then, the buyer has to rate the seller (positively or negatively). All sessions were conducted in May 2010 at the Behavioral and Experimental Economics Laboratory (BEELab) of the University of Florence, Italy. The 90 subjects (30 for each of the three treatments, 46 females and 44 males) were college students from various fields of study (mostly undergraduate students in business, 67%, or economics, 23%), who responded to an email (these were already in our mailing list of interested subjects) or received flyers advertising the experiment. The conversion rate between experimental points and Euros was 1.5 in Treatments 1 and 3 and 2 in Treatment 2 (this is because the expected average payoff was lower in this case and we wanted to have as similar saliency of the monetary incentive as possible). Participants worked on computers separated by partitions and were given written instructions (reproduced in Appendix A), which were read out loud by the lab staff and their comprehension was tested through a series of computerized questions. In addition, 3 practice rounds were played before starting with the 30 actual rounds. At the end of the session, each subject was privately paid in cash plus a 3€ show-up fee. The average payment per subject was €15.60 (with no appreciable differences in each treatment). The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

4. Experimental results

We firstly turn to the treatment effects. The related discussion is mainly based on a series of Mann-Whitney pairwise tests whose detailed results are reported in Table B2, Appendix B. Then, we further investigate the determinants of the feedback using regression analysis.

4.1 Treatment effects

Table 2 shows some descriptive statistics about the sellers.

Table 2. Sellers' descriptive statistics

	Treatment 1		**Treatment 2**		**Treatment 3**	
	Type A	**Type B**	**Type A**	**Type B**	**Type A**	**Type B**
Total transactions	457	443	463	437	451	449
Average transactions	20.8 (1.2)	19.3 (0.9)	21.1 (1.4)	19.0 (1.5)	20.5 (1.2)	19.5 (1.0)
Average buyer's payoff	37.2 (12.9)	18.2 (10.8)	31.1 (5.6)	10.0 (5.7)	33.6 (14.0)	22.3 (12.3)
Average lie	10.7 (5.8)	-10.6 (5.8)	9.5 (12.5)	-12.9 (13.0)	7.1 (9.7)	-6.7 (8.9)
Final positive ratings	83.0% (8.4%)	40.3% (8.9%)	93.4% (7.3%)	14.7% (9.4%)	77.0% (11.4%)	45.6% (8.1%)

Note: standard deviations in parenthesis.

The number of transactions pertaining to the two different types of sellers is similar across the treatments: this reflects the inverse relation between the sellers' probability of appearing in the triples presented to each buyer and the number of transactions.[8]

The third and fourth rows show the average payoff provided and lies told by types A and B, substantiating our definition of them as good and bad sellers respectively. The final row displays the positive feedback rate resulting from the treatments for the two types of sellers, and it allows us to evaluate the relative efficiency of the reputation mechanism in the three treatments.

Result 1. The observed behavior significantly differs from the socially optimal benchmark.

The experiment is characterized by a considerable fraction of transactions that are followed by suboptimal rating choices, in the sense that the rating does not reflect the type of seller correctly (24.48% over the three treatments). The largest proportion of suboptimal ratings is observed in Treatments 3 and 1.

Result 2. The final feedback profiles reflected the differences between the two types of sellers.

The differences between the average type A and type B sellers are remarkable in all treatments, which suggests that, once the sellers' profiles had been shaped, it was fairly simple for the buyers to detect the seller's type before the transaction. The second treatment presents the clearest profiles (i.e. closest to the ideal of 100% for types A and 0% for types B). Conversely, treatments 1 and 3 generated worse profiles for type A sellers and better profiles for type B sellers. Moreover, type B profiles in treatment 2 show a stark difference with respect to those of treatments 1 and 3, which are quite similar. Summing up:

Result 3. Different treatments resulted in different relative efficiency of the feedback system.

In treatment 2 the feedback profiles endogenously created clearly identified the sellers' types. In the other two treatments the system performed worse. Indeed it is worth remarking the following:

Result 3a. The efficiency of the feedback mechanism is lowest when both disconfirmation and payoff are confounding stimuli (i.e. in treatment 3).

We now turn to how the subjects left their feedback after the transactions. In particular we focus on the differences between the observations and the social optimum. Figures 4a-c show the fraction of positive ratings out of the total transactions for each possible value of the treatment-specific relevant signal (true value for treatments 1 and 3, payoff for treatment 2). The observed feedback behavior (dark line) is compared with the socially optimal behavior (bright line).

[8] Indeed such mechanism guaranteed that the advertised value (equal to 100 in all treatments) could be considered as a bona fide expected value: the average true value of the items sold was 100.2, 98.6, and 100.2 respectively in the three treatments.

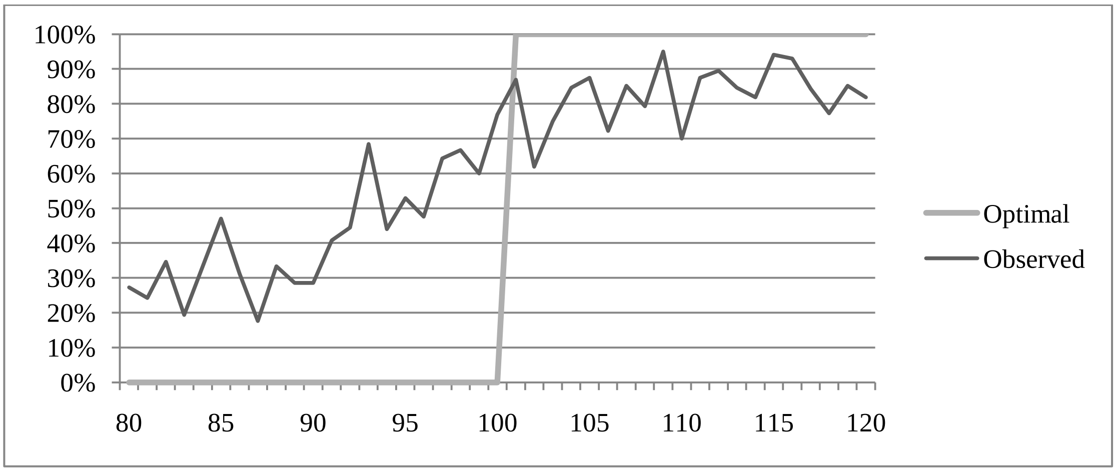

Figure 4a. % of positive ratings out of the total transactions for each obtained true value, tr. 1

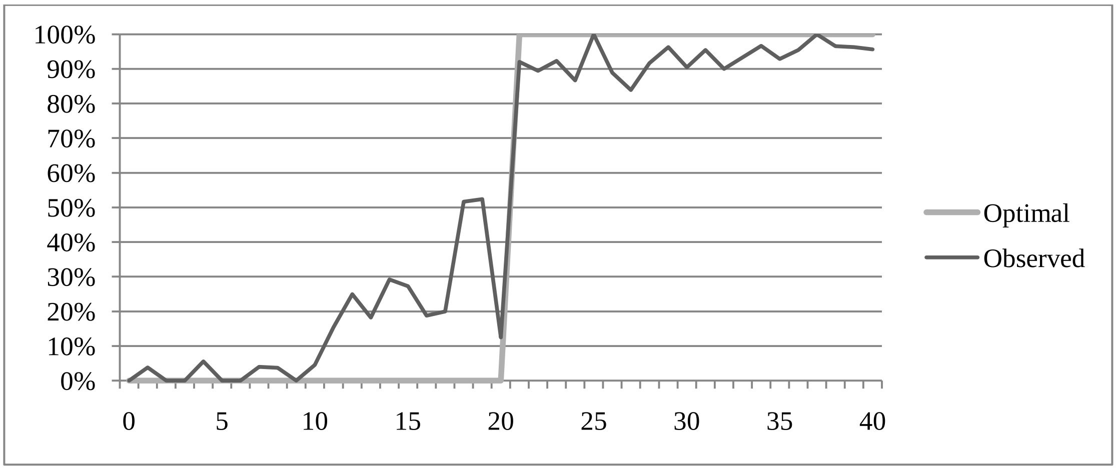

Figure 4b. % of positive ratings out of the total transactions for each realized payoff, tr. 2

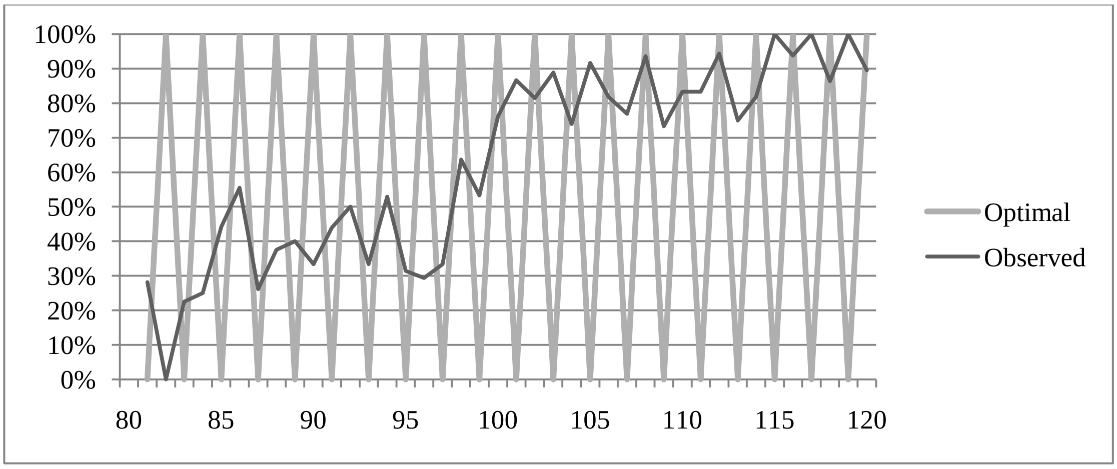

Figure 4c. % of positive ratings out of the total transactions for each obtained true value, tr. 3

Figure 4b for example shows that in treatment 2 the observed behavior was close to the social optimum: type A sellers almost always got a positive feedback, while types B received negative feedback when granting very low payoffs, but were rewarded with positive ratings when ensuring payoffs close to the threshold (payoff=21). In treatment 2 the reputation system worked significantly better than in the other treatments, in particular treatment 3 (Figure 4c), in which subjects largely ignored the even/odd signal, following instead some other behavioral rules.

It is also of interest to study the patterns of “wrong” ratings (i.e. positive feedback left to type B sellers, and negative feedback left to types A) which are depicted in Figure 5 as fractions over total transactions for each round. One feature is the following.

Result 4. A positive rating bias was observed.

Figure 5 shows that in all treatments there were significantly more wrong ratings for type B sellers than for type A ones.

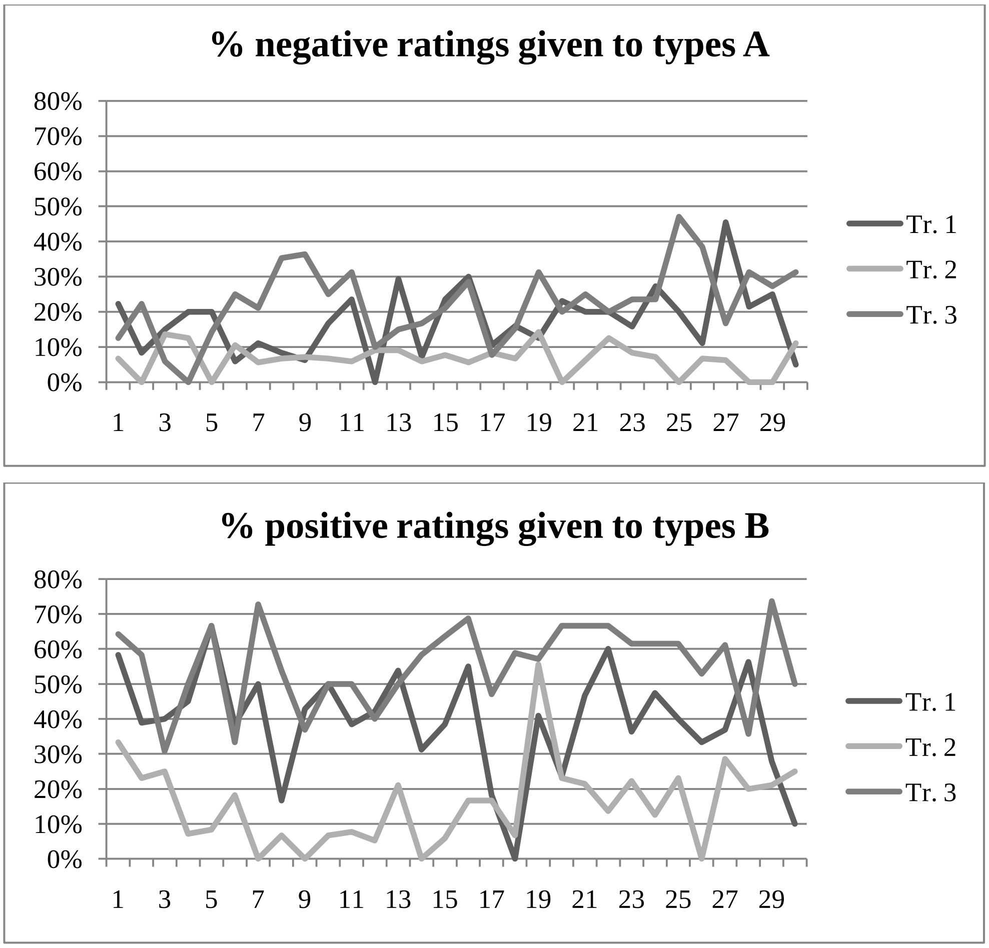

Figure 5. Fraction of “wrong” ratings per round

Notice that the fraction of wrong ratings is lowest for treatment 2, and highest for treatment 3, which was somehow already suggested by the final feedback profiles recorded in Table 2.[9] The tests reveal that there are significant differences in the wrong feedback patterns across treatments both for types A and for types B.

[9] Only in treatment 3 a type B seller ended up with a higher percentage of positive ratings with respect to a type A seller (65% vs. 55%).

We analyze in more details the fraction of wrong ratings, uncovering the different effects that the confounding stimuli exerted on the subjects' behavior. Table 3a-c present the numbers of ratings classified by lie told (whether hostile or pleasant) and payoff granted (more or less than 20) by the sellers, for the three treatments.

Table 3a-c. Number of ratings across payoffs and lies (absolute and percentages)

3a	Payoff 21-60		Payoff 0-20	
Tr. 1	pos	neg	pos	neg
Hostile Lie	118 57.84%	**86** **42.16%**	*61* *25.52%*	**178** **74.48%**
Pleasant Lie	**345** **86.25%**	*55* *13.75%*	**35** **61.40%**	22 38.60%

3b	Payoff 21-40		Payoff 0-20	
Tr. 2	pos	neg	pos	neg
Hostile Lie	**111** **90.98%**	11 9.02%	*42* *11.70%*	**317** **88.30%**
Pleasant Lie	**321** **94.13%**	*20* *5.87%*	24 30.77%	**54** **69.23%**

3c	Payoff 21-60		Payoff 0-20	
Tr. 3 – type A	pos	neg	pos	neg
Hostile Lie	**32** **55.17%**	26 44.83%	**24** **43.64%**	31 56.36%
Pleasant Lie	**274** **88.10%**	*37* *11.90%*	**18** **66.67%**	9 33.33%
	Payoff 21-60		Payoff 0-20	
Tr. 3 – type B	pos	neg	pos	neg
Hostile Lie	70 44.30%	**88** **55.70%**	*56* *28.43%*	**141** **71.57%**
Pleasant Lie	75 84.27%	**14** **15.73%**	4 80.00%	**1** **20.00%**

The bold cells contain the ratings optimally given by the subjects. The underlined cells contain the fraction of wrong ratings that may be attributed to the confounding stimuli that characterize the treatments. Finally, italics cells record the cases that cannot be explained in our simple framework.

Result 5. Most suboptimal ratings can be explained by the experimentally controlled confounding stimuli. Although in all treatments there are wrong ratings that elude our setup (in italics), they are always proportionally less than those (underlined) for which we can advance an explanation. These data also give more insight about the positive rating bias observed so far. In treatment 1, in 118 cases out of 204 (57.4%) a positive feedback was given to a type B granting a high payoff versus 22 cases out of 57 (38.6%) where a negative feedback was given to a type A seller that ensured a low payoff. Similarly, in treatment 2 the same emerges by comparing 30.8% of cases where a pleasant lie from a type B associated with a positive feedback, with 9.0% of cases where a type A seller was negatively rated after a hostile lie. For treatment 3 we must compare the fraction of confounding cases for type A sellers (66/451, 14.6%) with the ones for types B (149/449, 33.2%). We can deduce from this that the rating was positively biased in each and every treatment, independently of the nature of the confounding stimulus which we suppose to be at work.

We can also advance an explanation for the cases that elude the setup (in italics). Beside those who were induced to act sub-optimally by the presence of stimuli that clashed (as we discuss below in detail), but were nonetheless following some kind of logic, there was a fraction of subjects who were not willing to contribute meaningfully to the construction of the system itself, inserting a certain amount of noise in our data. Since taking part in fostering the reputation system was similar to a public good game (see Andreoni 1988, Fehr and Gachter 2000), contributing to which was compulsory and involved a small computational cost, one may classify such behavior as a kind of free-riding (in particular there where subjects who always left a positive feedback). Other subjects did bear the computational cost involved in recognizing the type but always reversed the rating with respect to type, which we may interpret as spiteful behavior (see Saijo and Nakamura, 1995) or downright confusion. Thus, the feedback profiles emerging from the dynamics of the experiment suffered from the behavior of such non-cooperators.

4.2 Regression analysis

In this subsection we present the results of our regression analysis on the determinants of feedback. Table 4 contains the probit estimations of the following model, which is a linear version of equation (1):

$$fb_{ijt} = \beta_0 + \beta_1(lie_{ijt}) + \beta_2(payoff_{it}) + \beta_3(fb_{jt-1}) + \varepsilon_t, \qquad (2)$$

where fb_{ijt} is a dichotomous variable whose value is 1 if buyer *i* left a positive feedback for seller *j*, and 0 if negative; lie_{ijt} is the difference between the true value and the advertised value; $payoff_{it}$ is the difference between the true value and the price; fb_{jt-1} is the percentage of past positive feedback ratings over the total effectuated by seller *j*; ε_t is an error term. The correlations among the variables are contained in Table B1a-c in Appendix B.

We estimate model (2) separately for each treatment (since in each of them the explanatory variable encoding the type of seller is different), augmenting the third estimate with a dichotomous variable (*evenodd*) taking the value of 1 when the true value of the item was even, indicating a type A seller.[10]

[10] Using a logit model as an alternative returned very similar results. We also estimated the model with fixed effects, again with little gain with respect to the baseline model.

Table 4. Random effects probit models, determinants of feedback

	Treatment 1		Treatment 2		Treatment 3	
	Coeff.	Change in prob.	Coeff.	Change in prob.	Coeff.	Change in prob.
payoff	*0.038*** (0.005)*	*0.204****	**0.113*** (0.011)**	**0.510*****	*0.029*** (0.005)*	*0.149****
lie	**-0.032*** (0.007)**	**-0.137*****	*-0.010* (0.006)*	*-0.062**	*-0.032*** (0.007)*	*-0.135****
fb_{t-1}	0.001 (0.002)	0.007	0.008*** (0.002)	0.130***	-0.001 (0.002)	-0.010
Constant	-0.611*** (0.231)		-2.424*** (0.217)		-0.523** (0.220)	
evenodd					**0.423*****	**0.152*****
N	900		900		900	
Pearson χ^2	290.68 (df = 3)		721.56 (df = 3)		271.18 (df = 4)	
Pseudo R^2	0.25		0.58		0.23	
% correctly predicted	72.00		88.56		71.78	

Note: *, **, *** significant at 10, 5 and 1% respectively; standard errors in parenthesis. Boldface indicates that the variable signals the type, italics indicates the variable encodes a confounding stimulus. The change in probability is the change in predicted probability of giving positive feedback for an increase of 1 standard deviation in each independent variable (for the *evenodd* dummy the change from 0 to 1 is considered).

Result 6. Both *lie* and *payoff* significantly contribute to the feedback behavior. The same does not hold for the seller's past cumulated feedback.

The relevance of the variable encoding the type of seller and of the confounding variables differs widely across treatments. In treatment 1 changes in *payoff*, which was confounding, had a larger impact on the probability of leaving a positive feedback than changes in the optimally signaling variable, *lie* (a one standard deviation increase changes the probability of giving a positive feedback by .20 for *payoff* and -.14 for *lie*). In treatment 2, the situation is reversed: a one standard deviation rise in *payoff* increases the probability of giving a positive feedback by .51, versus an estimated -.06 for a similar change in *lie*. In treatment 3, the *evenodd* variable has a comparable bearing to that of *lie* and *payoff*, both of which were confounding. In particular, the fact that the true value of the delivered item is an even number rises the probability of giving a positive feedback by .15, as approximately does (in absolute value) a standard deviation rise in either *lie* or *payoff*. The importance of past cumulated feedback is low: coefficients and marginal effects are either not significantly different from zero (treatments 1 and 3) or quantitatively small (treatment 2). Finally notice that in all cases the models fit the observed behavior of the subjects fairly well, especially in the second treatment where more than 88% of the cases are correctly predicted.

5. Conclusions

This paper studies the determinants of positive and negative feedback on sellers within an experimental trading environment in which a reputation system is endogenously fed by post-transactions buyers' ratings. The efficiency of the resulting feedback system is analyzed in detail.

The economic payoff of the transaction emerges as the leading factor driving the direction of the ratings. Therefore, obtaining a sufficient payoff stacked the odds in favor of ending up rating the seller positively, possibly even when there was unambiguous evidence that the seller was a bad one. Similarly, but to a lesser extent, the reverse also happened, as a good seller received a negative feedback after a transaction in which the payoff for the buyer was low due to a mishap.

One practical implication of these results is that they account for the possible existence of a propensity of sellers on online marketplaces to exceed their marketing efforts to hype up the descriptions of goods on sale. Documented examples include the case of online markets for cards, stamps or other collectibles, in which many items happen to be sold as "mint" when their actual quality is lower and where nonetheless most transactions normally end up receiving a positive feedback, especially when the selling price is appropriate for the goods' actual worth (see Jin and Kato, 2006). Such policy, while not boosting profits per se, has the consequence of easing sales while not compromising reputation.

For what concerns the practical design implications, our findings support the inclusion of a measure of the quality price ratio obtained in the purchase within the detailed seller ratings available to buyers beside the traditional feedback score and percentage that usually form the backbone of electronic reputation. In particular this may be appropriate when the seller effectively sets the price. This might help subsequent customers seeing through the usual mass of uniformly excellent feedback profiles typically encountered, by expanding the available information.

An aggregate implication arising from our data is that, overall, our experimental ratings exhibited a positive rating bias: buyers were more likely to rate positively bad sellers than to rate negatively good sellers. Our experiment suggests that this may be the result of sellers granting a sufficient payoff to buyers, rather than keeping a spotless behavior. That is, perhaps not all the retailers are "objectively" high quality sellers, but they may behave in such a way to make buyers satisfied enough to guarantee a positive feedback, or at least their silence (see Dellarocas and Wood 2008). To what extent such misunderstanding can be held responsible to contribute to the failure of the online reputation systems to keep fraud at bay (see Gavish and Tucci 2008) warrants more context-specific research.

The analysis carried out in this article leads naturally to additional research questions, regarding the robustness of our results to changes in various institutional and parametric assumptions. In particular, it could be worth investigating the effects of losses. Also, it would be interesting to introduce an auction mechanism in the allocation of goods and see how this modifies behavior and in which direction. This would make it possible to ascertain whether and how, moving to a situation in which the buyer, to some degree, makes the price and therefore affects the quality price ratio, modifies buyers' behavior.

References

Akerlof, G. 1970. The Market for ‘Lemons’: Quality Under Uncertainty and the Market Mechanism. *Quarterly Journal of Economics* 84 488-500.

Anderson, E.W. 1998. Customer satisfaction and word of mouth. *Journal of Service Research* 1(1) 5-17.

Anderson, E.W., M.W. Sullivan. 1993. The antecedents and consequences of customer satisfaction for firms. *Marketing Science* 12(2) 125-143.

Andreoni, J. 1988. Why Free Ride? Strategies and Learning in Public Goods Experiments. *Journal of Public Economics* 37(3) 291-304.

Asch, S.E. 1956. Studies of independence and conformity: 1, a minority of one against a unanimous majority. *Psychological Monographs* 70(9) No. 416 1-70.

Ba, S., P. Pavlou. 2002. Evidence of the Effect of Trust Building Technology in Electronic Markets: Price Premiums and Buyer Behavior. *MIS Quarterly* 26(3) 243-268.

Bolton, G., E. Katok, A. Ockenfels. 2004. How Effective are Online Reputation Mechanisms? An Experimental Investigation. *Management Science* 50 1587-1602.

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

Dellarocas, C., C. Wood. 2008. The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias. *Management Science* 54(3) 460-476.

Fehr, E., S. Gachter. 2000. Cooperation and Punishment in Public Goods Experiments. *The American Economic Review* 90(4) 980-994.

Fischbacher, U. 2007. z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10(2) 171-178.

Gavish, B., C.L. Tucci. 2008. Reducing internet auction fraud. *Communications of the ACM* **51(5)** 89-97.

Gazzale, R.S., T. Khopkar. 2011. Remain silent and ye shall suffer: seller exploitation of reticent buyers in an experimental reputation system. Experimental Economics DOI 10.1007/s10683-010-9267-z.

Gefen, D., E. Karahanna, D. Straub. 2003. Trust and TAM in Online Shopping: An Integrated Model. *MIS Quarterly* **27(1)** 51-90.

Gregg D.G., J.E. Scott. 2006. The Role of Reputation Systems in Reducing On-Line Auction Fraud. *International Journal of Electronic Commerce* 10(3) 95-120.

Houser, D., J. Wooders. 2006. Reputation in Auctions: Theory, and Evidence from eBay. *Journal of Economics and Management Strategy* 15(2) 353-369.

Inman, J.J., J.S. Dyer, J.J. Jianmin. 1997. A generalized utility model of disappointment and regret effects on post-choice valuation. *Marketing Science* 16(2) 97-111.

Jarvenpaa, S., N. Tractinsky, M. Vitale. 2000. Consumer Trust in an Internet Store. *Information Technology and Management* 1 45-71.

Jin, G.Z., A. Kato. 2006. Price, Quality, and Reputation: Evidence from an Online Field Experiment. *The RAND Journal of Economics* 37(4) 983-1004.

Josang, A., R. Ismail, C. Boyd. 2006. A Survey of Trust and Reputation Systems for Online Service Provision. *Decision Support Systems* 43 618-644.

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

Klein, T.J., C. Lambertz, G. Spagnolo, K.O. Stahl. 2009. The actual structure of eBay's feedback mechanism and early evidence on the effects of recent changes. *International Journal of Electronic Business* 7(3) 301-320.

Kohen, D. 2003. The Nature of and Conditions for Online Trust. *Journal of Business Ethics* 43(3) 3-19.

Li, S., K. Srinivasan, B. Sun. 2009. Internet Auction Features as Quality Signals. *Journal of Marketing* 73 75-92.

Livingston, J. 2005. How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions. *The Review of Economics and Statistics* 87(3) 453-465.

Lucking-Reiley, D., D. Bryan, N. Prasad, D. Reeves. 2007. Pennies from eBay: The Determinants of Price in Online Auctions. *Journal of Industrial Economics* 55(2) 223-233.

McDonalds, C., C. Slawson. 2002. Reputation in an Internet Auction Market. *Economic Inquiry* 40(3) 633-650.

Melnik, M., J. Alm. 2002. Does a Seller’s Ecommerce Reputation Matter? Evidence from eBay Auctions. *The Journal of Industrial Economics* 50(3) 337-349.

Moon, J.W., Y.G. Kim. 2001. Extending the TAM for a World-Wide-Web Context. *Information & Management* 38 217-230.

Novemsky, N., D. Kahneman. 2005. The Boundaries of Loss Aversion. *Journal of Marketing Research* 42(2) 119-128.

Pavlou, P. 2003. Consumer Acceptance of Electronic Commerce: Integrating Trust and Risk with the Technology Acceptance Model. *International Journal of Electronic Commerce* 7(3) 69-103.

Resnick, P., R. Zeckhauser. 2002. Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay’s Reputation System. *Advances in Applied Microeconomics: A Research Annual* 11 127-157.

Rust, R.T., J.J. Inman, A. Zahorik. 1999. What you don’t know about customer-perceived quality: The role of customer expectations distributions. *Marketing Science* 18(1) 77-92.

Saijo, T., H. Nakamura. 1995. The “spite” dilemma in voluntary contribution mechanism experiments. *Journal of Conflict Resolution* 39 535-560.

Söderlund, M. 1998. Customer satisfaction and its consequences on customer behaviour revisited. *International Journal of Service Industry Management* 9(2) 169-188.

Standifird, S. 2001. Reputation and e-commerce: eBay Auctions and the Asymmetrical Impact of Positive and Negative Ratings. *Journal of Management* 27 279-295.

Utz, S., U. Matzat, C. Snijders. 2009. Online Reputation Systems: The Effects of Feedback Comments and Reactions on Building and Rebuilding Trust in Online Auctions. *International Journal of Electronic Commerce* 13(3) 95-118.

Yang, J., X. Hu, H. Zhang. 2007. Effects of a Reputation Feedback System on an Online Consumer-to-Consumer Auction Market. *Decision Support Systems* 44 93-105.

Zhang, J. 2006. The Roles of Players and Reputation: Evidence from eBay Online Auctions. *Decision Support Systems* 42 1800-1818.

Appendix A – Subjects' instructions

Below are the written instructions that were given to subjects in the third (even/odd) treatment. Instructions for the other treatments were parallel, the only differences being the description of the types of sellers (different distributions of the true values delivered after the transaction). The sessions were held on 05/04/2010 2 PM (treatment 1), 05/06/2010 2PM (treatment 2) and 05/11/2010 2PM (treatment 3).

Experiment description. The purpose of this session is to study how people make decisions when facing uncertainty/ambiguity.

The session will last 30 rounds. Throughout the experiment, values, prices and payoffs are expressed in a fictional unit of account, called "points". At the end, the points that you will obtain will be converted in € according to the following conversion rate:

(total payoff in points) x 1.5 = total payoff in € cents (e.g. 900 points equal 13.50€)

Every round you must buy a fictional item that you don't directly observe. The value advertised by the seller through a description is 100 points; it does not necessarily coincide with the true value of the item, which will be revealed only after the transaction is concluded (as in the case of a mail order and of online shopping).

Each round you will face three different sellers, picked randomly from a total of 45. Sellers are computer-generated and are different one from another. Their behavior is determined by the software following the logic described below: they don't correspond to any real person. You will choose from which one of the three to buy the item. For each seller the following information will be given:

1) the price at which to buy the item (picked randomly from the following values: 60, 65, 70, 75, 80, 85, 90, 95, 100);

2) the seller's profile (see the paragraph "The sellers' profile").

Once you choose, you will discover the true value of the item that you bought. Each round, your payoff is equal to the difference between the true value and the price paid. Example:

price paid = 90; true value = 112; payoff of the round = (112 – 90) = 22 points.

At the end of each round you will be asked to leave a feedback, "positive" or "negative", about the seller. The ratings expressed by you and the other participants will be used to update the sellers' profiles (for the sellers that had sold items during the round). This information system may help you choosing the sellers in the following rounds.

The sellers' profile. Every seller is characterized by a profile containing the following information:

- the name of the seller, composed by 4 alphanumeric characters (e.g. XY95);

- the number of the transactions effected so far;
- the number of positive and negative ratings obtained;
- the percentage of positive ratings, calculated in the following way: number of positive ratings divided by the number of transactions effected.

Different types of sellers. Sellers are not all equal. They differ in terms of quality (true value) of the item that they deliver after your choice.

There are two types of sellers, type "A" and type "B".

Type "A" sellers deliver items whose true value is an **even** integer picked randomly in the interval 80-120. The probability with which the random pick is done is described by the following figure:

(The figure that appeared here is the one of the upper panel - Figure 4 in the text of the paper)

Thus, type "A" sellers deliver items whose true value is 80 with a lower probability with respect to items with a true value of 82; accordingly, this last probability is lower than the one of delivering an item whose true value is 84, and so on until 120 (the true value associated with the highest probability). Note that type "A" sellers deliver, on average, 3 times out of 4 items whose true value is equal or higher than the advertised value, that is, lies between 100 and 120.

Type "B" sellers deliver items whose true value is an **odd** integer picked randomly in the interval 81-119. The probability with which the random pick is done is described by the following figure:

(The figure that appeared here is the one of the lower panel - Figure 4 in the text of the paper)

Thus, type "B" sellers deliver items whose true value is 81 with a higher probability with respect to items with a true value of 83; accordingly, this last probability is higher than the one of delivering an item whose true value is 85, and so on until 119 (the true value associated with the lowest probability). Note that type "A" sellers deliver, on average, 3 times out of 4 items whose true value is lower than the advertised value, that is, lies between 81 and 99.

Sellers are half type "A" and half type "B", and you will choose among three of them each round.

Note that:

1) the price does not signal the type of seller in any respect;

2) the true value may be lower than the price paid: in this case a "satisfied or reimbursed" clause is at work, that is, you will be given the price paid back and your payoff in this round will be zero;

3) you are not competing with the other participants.

Money earnings. You will be paid your earnings from all the rounds of the session (number of points multiplied by 1.5 plus a 3€ show up fee) in cash.

Comprehension questions and practice rounds. When the monitor gives the OK, you first have to answer some questions regarding the comprehension of the information given so far. Then, you have to play some practice rounds. During these rounds you can familiarize with the choices that you have to make during the paid session (that is, choice of the seller from which to buy and rating of the seller after the transaction). Note: during the practice rounds, the sellers' names will be "Seller 1", "Seller 2"…, differently from the alphanumerical names that will be used during the paid session.

Appendix B – Correlations and pairwise Mann-Whitney tests

Table B1a – Cross-Correlations – Treatment 1

	Pos. Fb.	Payoff	Lie	Past Feedback
Pos. Fb.	1.00			
Payoff	0.48	1.00		
Lie	-0.46	-0.74	1.00	
Past Feedback	0.24	0.33	-0.55	1.00

Note: Pos. Fb. = 0 if negative, 1 if positive (dependent variable).

Table B1b – Cross-Correlations – Treatment 2

	Pos. Fb.	Payoff	Lie	Past Feedback
Pos. Fb.	1.00			
Payoff	0.77	1.00		
Lie	-0.59	-0.74	1.00	
Past Feedback	0.66	0.74	-0.59	1.00

Note: Pos. Fb. = 0 if negative, 1 if positive (dependent variable).

Table B1c – Cross-Correlations – Treatment 3

	Evenodd	Pos. Fb.	Payoff	Lie	Past Feedback
Evenodd	1				
Pos. Fb.	0.32	1.00			
Payoff	0.39	0.43	1.00		
Lie	-0.60	-0.47	-0.75	1.00	
Past Feedback	0.48	0.14	0.16	-0.28	1.00

Note: Pos. Fb. = 0 if negative, 1 if positive (dependent variable). Evenodd = 0 if odd true value, 1 if even.

Table B2. Mann-Whitney tests

Variable of interest	**Tr. 1 vs Tr. 2**	**Tr. 1 vs Tr. 3**	**Tr. 2 vs Tr. 3**	**Tr. 1**	**Tr. 2**	**Tr. 3**
Final no. of transactions, type A sellers	-0.589 (0.5559)	0.742 (0.4580)	1.278 (0.2012)			
Final no. of transactions, type B sellers	0.91 1 (0.3620)	-0.622 (0.5336)	-1.401 (0.1613)			
Final % of pos. feedback, type A vs. type B				-5.752*** (0.0000)	-5.771*** (0.0000)	-5.649*** (0.0000)
Final % of pos. feedback, type A sellers	-3.765*** (0.0002)	2.115*** (0.0344)	4.394*** (0.0000)			
Final % of pos. feedback, type B sellers	5.446*** (0.0000)	-1.874* (0.0609)	-5.730*** (0.0000)			
Wrong feedback: type A vs. type B				-5.088*** (0.0000)	-3.184*** (0.0015)	-6.384*** (0.0000)
Wrong feedback: pos. fb. to type B sellers	4.614*** (0.0000)	-2.169** (0.0301)	-5.656*** (0.0000)			
Wrong feedback: neg. fb. to type A sellers	5.118*** (0.0000)	-4.034*** (0.0001)	-6.444*** (0.0000)			

Note: **, *** significant at 5 and 1% respectively. P-values in parenthesis. Null hypothesis: there is symmetry between the two populations.