

U N I V E R S I T Y O F C O L O G N E
W O R K I N G P A P E R S E R I E S I N E C O N O M I C S

**ENGINEERING TRUST - RECIPROCITY IN THE
PRODUCTION OF REPUTATION INFORMATION**

**GARY BOLTON
BEN GREINER
AXEL OCKENFELS**

Department of Economics
University of Cologne
Albertus-Magnus-Platz
D-50923 Köln
Germany

<http://www.wiso.uni-koeln.de>

ENGINEERING TRUST

- RECIPROCITY IN THE PRODUCTION OF REPUTATION INFORMATION -

GARY BOLTON, BEN GREINER, AND AXEL OCKENFELS

22 January 2011

Abstract. Reciprocity in feedback giving distorts the production and content of reputation information in a market, hampering trust and trade efficiency. Guided by feedback patterns observed on eBay and other platforms we run laboratory experiments to investigate how reciprocity can be managed by changes in the way feedback information flows through the system, leading to more accurate reputation information, more trust and more efficient trade. We discuss the implications for theory building and for managing the redesign of market trust systems.

Keywords: market design, reputation, trust, reciprocity, eBay

JEL classification: C73, C9, D02, L14

Financial support from the German Science Foundation (DFG) and U.S. National Science Foundation (NSF) is gratefully acknowledged. We thank Brian Burke, Debbie Hofmeyr, Leland Peterson and the rest of eBay's Trust&Safety team for their cooperation in this project. We provided eBay with advice on improving their feedback system under an agreement that permitted eBay site data be used for academic publication. We also thank seminar participants at Berlin, Brisbane, Cologne, Copenhagen, Graz, Harvard, Indiana, Luxembourg, Max Planck Jena, Michigan, Royal Holloway London, Melbourne, Nürnberg, Santa Barbara, and Sydney for helpful comments. We are indebted to Felix Lamouroux, Karin Ruetz, and Dietmar Ilsen for excellent research assistance and help with the collection of data.

Bolton. Pennsylvania State University, Smeal College of Business, University Park, PA 16802, Tel: 1 (814) 865 0611, Fax: 1 (814) 865 6284, e-mail: gbolton at psu.edu.

Greiner. University of New South Wales, School of Economics, Sydney, NSW 2052, Australia, Tel: +61 2 9385 9701, Fax: +61 2 9313 6337, e-mail: bgreiner at unsw.edu.au.

Ockenfels. University of Cologne, Department of Economics, Albertus-Magnus-Platz, D-50923 Köln, Germany, Tel: +49/221/470-5761, Fax: +49/221/470-5068, e-mail: ockenfels at uni-koeln.de.

I. Introduction

This paper reports on the repair of an Internet market trust mechanism. While all markets require some minimum amount of trust (Akerlof, 1970), it is a particular challenge for Internet markets, where trading is typically anonymous, geographically dispersed, and executed sequentially. To incentivize trustworthiness, Internet markets commonly employ reputation-based ‘feedback systems’ that enable traders to publicly post information about past transaction partners. Online markets with this kind of feedback system include eBay.com, Amazon.com, and RentACoder.com, among many others. For these markets, feedback systems with their large databases of transaction histories are a core asset, crucial for market efficiency and user loyalty.

Below we will see, based on new data and reports of other researchers, that there is a strong reciprocal pattern in the content, timing and quantity of feedback information given in the eBay marketplace. The propensity for reciprocity may reduce the informativeness of the feedback given and so hamper market efficiency. It turns out that the institutional trigger for this behavior is the rules governing feedback timing and observability.

Our approach to fixing this problem combines behavioral economics with an engineering perspective: The objective is to discover how scientifically developed market design ideas will likely perform within an operating marketplace *prior* to implementation, thereby reducing the risk of costly mistakes due to unforeseen or underestimated circumstances. Towards this end, we study two alternative redesigns, both proposed in the literature as methods to diminish reciprocal feedback behavior. Using a specifically designed laboratory experiment, complemented by additional evidence from the field, we evaluate the robustness of the redesigns with respect to complications embedded in the existing market feedback mechanism and with respect to negative side effects in other parts of the market system. Our analysis guided eBay in its decision to change the reputation system. We present preliminary data on how the post-fix system works in practice.

Engineering studies put science to a prescriptive test. In the present case, simple theory implies that a reputation system that elicits accurate and complete feedback information can promote trust and cooperation among selfish traders even in such adverse environments as online market platforms (e.g., Wilson, 1985, Milgrom et al., 1990). So there is theoretical reason to believe that a properly designed feedback system can effectively facilitate trade. At the same time, our data take us further down the causation chain than received theory presently goes, to gaming in the production of reputation information. In essence, reputation builders retaliate for a negative review, thereby inhibiting the provision of negative reviews in the first place. The resulting bias in reputation information then works its way up the chain, ultimately diminishing market efficiency.¹ This complication challenges the usefulness of existing concepts of reputation building that

¹ Another problem of distorted feedback information arises in the image scoring game, which has first been studied by Nowak and Sigmund (1998). In each period of the game, players are paired with one given the chance to take a costly

assume away endogenous feedback production. That is, market design can help identify the gaps of existing concepts, lead to new research questions, and establish the effectiveness of proposed solutions (see Ostrom 1990 and Roth 2002 for pioneering work along these lines).²

However, any attempt to fix the problem by changing the institutional rules involves two complexities. First, it is not clear how responsive the system would be: In order to be economically effective, the new system need evoke subtle, strategically motivated changes in the behavior of the traders, both regarding feedback provision and trade conduct, as the information flows through the market. Second, changing the feedback rules risks undesirable side effects. In particular, reciprocal feedback has positive, as well as negative consequences. Most critically, reciprocity appears important to getting (legitimately) satisfactory trades reported; eliminating this kind of reciprocity might lead to a system that over-reports, rather than under-reports, negative outcomes. In addition, market design cannot ignore path dependencies. eBay's feedback system is synchronized with other parts of the market platform, such as eBay's conflict resolution system, so that significant changes in one part would often entail major changes in other parts. Also, exchanging the old feedback system by a 'theoretically better' system might lead to millions of eBay traders losing their old reputations in the transition to the new system. Thus, in market design, small changes – if effective – are often preferred to large changes.³

With these considerations in mind, our laboratory study examines two alternative proposals (described in Section II).⁴ Descriptive data from other Internet markets that have feedback systems with features similar to those proposed provide answers to some of our questions (Section III). But not all of them: Behavioral and institutional differences across the markets create substantial ambiguity; one proposal, in particular, has major features not shared with any existing market. Also, we lack field data on the underlying cost and preference parameters in the markets, and so cannot easily measure how feedback systems affect market efficiency. To narrow the uncertainty, we crafted a test bed experiment designed to capture the theoretically relevant aspects

action that helps the other. Cooperating in this manner is socially efficient, but the only way to monitor free riding is through the image score (reputation) which in this game comes to an accounting of a player's past helping actions. Here, too, agents may not want to punish cheaters, because this risks spoiling one's own reputation (see Bolton et al., 2005, for laboratory evidence).

² Roth (2002, p. 1341) argues the need for a literature "to further the design and maintenance of markets and other economic institutions" and provides examples of how this literature can help shape new questions in theory. To date, the market design literature has focused mostly on allocation mechanisms such as auctions and matching. Roth (2008) reviews the literature on matching markets, Milgrom (2004) the literature on auction markets, and Greiner et al. (forthcoming) the literature specific to Internet auctions. Regarding market design, the work probably closest to ours is Chen et al. (2010), who show in the context of a field experiment on MovieLens that personalized social information flows can increase the level and quality of public good contributions (namely the number of movie ratings submitted by a user).

³ For a further discussion of this problem see, for example, Niederle and Roth (2005).

⁴ A number of others proposals were considered but discarded relatively quickly in favor of the two discussed here.

of behavior and institutional changes (Section IV).⁵ In combination with the field observations, the lab data provide a robust picture of how reciprocity flows through information channels. We present preliminary data on how the eventually implemented new field system performs (Section V).

Our study adds to previous evidence that there is a great deal of reciprocal behavior in the production of feedback information, and that this behavior has both positive (participation) and negative (biasing) effects on the reputation system. The engineering dimension of our study takes us an important step further: It illustrates that carefully targeted changes to the feedback system can set off an endogenous shift in the market, leading to greater trust and trustworthiness, and ultimately more efficient trade. The lessons learned in this study extend far beyond the scope of eBay's feedback system: other reputation-based systems, both online and offline, are open to similar retaliation issues (ex., financial rating services, employee job assessments, word-of-mouth about colleagues), so the phenomenon is likely to be found in many markets and social environments involving reputation building, and the fixes and insights discussed here may be helpful to tackle problems elsewhere. We discuss the implications for theory building and for managing the redesign of market trust systems in Section VI.

II. The feedback problem and two proposals to fix it

In this section we first review eBay's conventional feedback system (Subsection II.1). We then examine evidence, from new data as well as from the work of other researchers, for a reciprocal pattern in feedback giving and for the role of the rules that govern feedback giving (Subsection II.2). An important point will be that reciprocal behavior appears to have good as well as bad consequences for the system.⁶ We then discuss two proposals put forward to mitigate the problem, which will be the subject of our laboratory test (Subsection II.3).

⁵ Test bed experiments to get insight into how a market redesign will work have been done in relation to allocation mechanisms; for example, Grether, Isaac, and Plott (1981), Kagel and Roth (2000), Chen (2005), Kwasnica, Ledyard, Porter, and DeMartini (2005), Chen and Sönmez (2006) and Brunner, Goeree, Holt, and Ledyard (forthcoming).

⁶ That said, many (but not all) studies find that feedback has positive value for traders as indicated by positive correlations between the feedback score of a seller and the revenue and the probability of sale. See, for example, Bajari and Hortaçsu (2003, 2004), Ba and Pavlou (2002), Cabral and Hortaçsu (2010), Dellarocas (2004), Dewan and Hsu (2001), Eaton (2007), Ederington and Dewally (2006), Houser and Wooders (2005), Jin and Kato (forthcoming), Kalyanam and McIntyre (2001), Livingston (2005), Livingston and Evans (2004), Lucking-Reiley, Bryan, Prasad, and Reeves (2007), McDonald and Slawson (2002), Melnik and Alm (2002), Ockenfels (2003), Resnick and Zeckhauser (2002), and Resnick, Zeckhauser, Swanson, and Lockwood (2006). See Ba and Pavlou (2002), Bolton, Katok, and Ockenfels (2004, 2005), and Bolton and Ockenfels (2009) for laboratory evidence. Further related experimental evidence is provided in Dulleck, Kerschbamer and Sutter (forthcoming), who investigate potentially efficiency-enhancing mechanisms in large experimental credence goods markets, which are – like eBay – also characterized by asymmetric information between sellers and consumers, and Sutter, Haigner and Kocher (forthcoming), who find large and positive effects on cooperation in an experimental public goods game if group members can endogenously determine its institutional design. Lewis (2010) studies endogenous product disclosure choices of sellers of used cars on eBay as a complementary mechanism contributing to overcome problems of asymmetric information in the market place.

II.1 eBay's conventional feedback system

eBay facilitates trade in the form of auctions and posted offers in over thirty countries.⁷ After each eBay transaction, both the buyer and the seller are invited to give feedback on each other. Until spring 2007, when eBay changed the system, only “conventional” feedback could be left. In the conventional feedback system, a trader can rate a transaction positive, neutral, or negative (along with a text comment). Submitted feedback is immediately posted and available to all traders. Conventional feedback ratings can be removed from the site only by court ruling, or if the buyer did not pay, or if both transaction partners mutually agree to withdrawal.⁸

The most common summary measure of an eBay trader's feedback history is the *feedback score*, equal to the difference between the number of positive and negative feedbacks from unique eBay traders (neutral scores are ignored). Each trader's feedback score is provided on the site. An important advantage of the feedback score is that it incorporates a reliability measure (experience) in the measure of trustworthiness. The feedback score is also the most commonly used measure of feedback history in research analyses of eBay data.⁹

II.2 Reciprocal feedback

Feedback information is largely a public good, helping other traders to manage the risks involved in trusting unknown transaction partners. Yet our data finds that about 70% of the traders leave feedback (a number consistent with previous research).¹⁰ In this subsection, we examine evidence that reciprocity plays a role in the giving as well as timing and content of eBay feedback. In the following, the null hypothesis is always that feedback is given independently, whereas the alternative hypothesis states that feedback is given conditionally, following a reciprocal pattern.¹¹ The analysis is based on 700,000 completed eBay transactions taken from seven countries and six categories in 2006/07.¹²

⁷ In 2007, 84 million users bought or sold \$60 billion in goods on eBay platforms.

⁸ eBay's old feedback system was the product of an 11 year evolutionary process. In its first version, introduced in 1996, feedback was not bound to mutual transactions: every community member could give an opinion about every other community member. In 1999/2000 the ability to submit non-transaction related feedback was removed. The percentage of positive feedback was introduced in 2003, and in 2004 the procedure of mutual feedback withdrawal was added. Since 2005, feedback submitted by eBay users leaving the platform shortly thereafter or not participating in ‘issue resolution processes’ is made ineffective, and members who want to leave neutral or negative feedback must go through a tutorial before being able to do so. Since spring 2007 a new system was introduced, as described in Section V. In 2008, again new features were implemented, which are analyzed in Bolton, Greiner and Ockenfels (2010).

⁹ Another common measure is the ‘percentage positive’ equal to the share of positive and negative feedbacks that is positive. For our data, which measure is used makes little difference; we mostly report results using the feedback score.

¹⁰ The number varies somewhat across categories and countries. Resnick and Zeckhauser (2002) found that buyers gave feedback in 51.7% of the cases, and sellers in 60.6%. Cabral and Hortaçsu (forthcoming) report a feedback frequency from buyer to seller in 2002/03 of 40.7% in 1,053 auctions of coins, notebooks and Beanie Babies. In their 2002 dataset of 51,062 completed rare coin auctions on eBay, Dellarocas and Wood (2008) observed feedback frequencies of 67.8% for buyers and 77.5% for sellers.

¹¹ Others recognized the strategic interdependency of feedback before, including Resnick, Zeckhauser, Friedman, and Kuwabara (2000), Resnick and Zeckhauser (2002), and, more recently, Klein, Lambertz, Spagnolo, and Stahl (2007) and Dellarocas and Wood (2008). In this and parts of the following sections, we complement this literature from an engineering perspective, combining data sets that follow a large set of eBay transactions from their posting until the end

TABLE 1: FEEDBACK GIVING AND CONTENT, CONDITIONAL PROBABILITIES AND CORRELATIONS

Feedback giving probability	Partner did not yet give FB		Partner gave FB already	
Buyer	68.4%		74.1%	
Seller	51.4%		87.4%	

Kendall's tau correlations between seller's and buyer's feedback									
Country	FB content correlation						FB giving correlation		
	All cases		Buyer gave		Seller gave				
	N	tau	N	Tau	N	tau	N	tau	
All	458,249	0.710	139,772	0.348	318,477	0.884	725,735	0.693	
Australia	20,928	0.746	6,040	0.340	14,888	0.928	31,990	0.752	
Belgium	8,474	0.724	3,097	0.464	5,377	0.880	12,301	0.684	
France	24,933	0.727	8,095	0.423	16,838	0.883	39,104	0.703	
Germany	133,957	0.656	45,836	0.331	88,121	0.840	192,565	0.644	
Poland	457	1.000	172	-	285	1.000	1,134	0.783	
U.K.	93,266	0.694	31,316	0.379	61,950	0.875	143,877	0.692	
U.S.	176,009	0.746	45,133	0.313	130,876	0.911	302,213	0.701	

Notes: Observations where feedback was eventually withdrawn are not included in correlations. In the cell marked with “-“, the standard deviation is zero. All other correlations are highly significant.

Feedback giving. If feedback were given independently among trading partners, one would expect the percentage of time both partners give feedback to be $70\% \times 70\% = 49\%$. Yet mutual feedback is given much more often, about 64% of the time. The reason is evident from the top rows of Table 1: Both buyers and sellers are more likely to provide feedback when the transaction partner has given feedback first. The effect is stronger for sellers than for buyers; when a buyer gives feedback, the seller leaves feedback 87.4% of the time, versus 51.4% when the buyer does not leave feedback (in a moment we will see that sellers sometimes have an incentive to wait). A common buying experience on eBay, after a transaction has gone smoothly, is to receive a note from the seller saying he gave you positive feedback and asking you to provide feedback, or saying that he would give you feedback once you left feedback on him (playing or initiating a kind of “trust

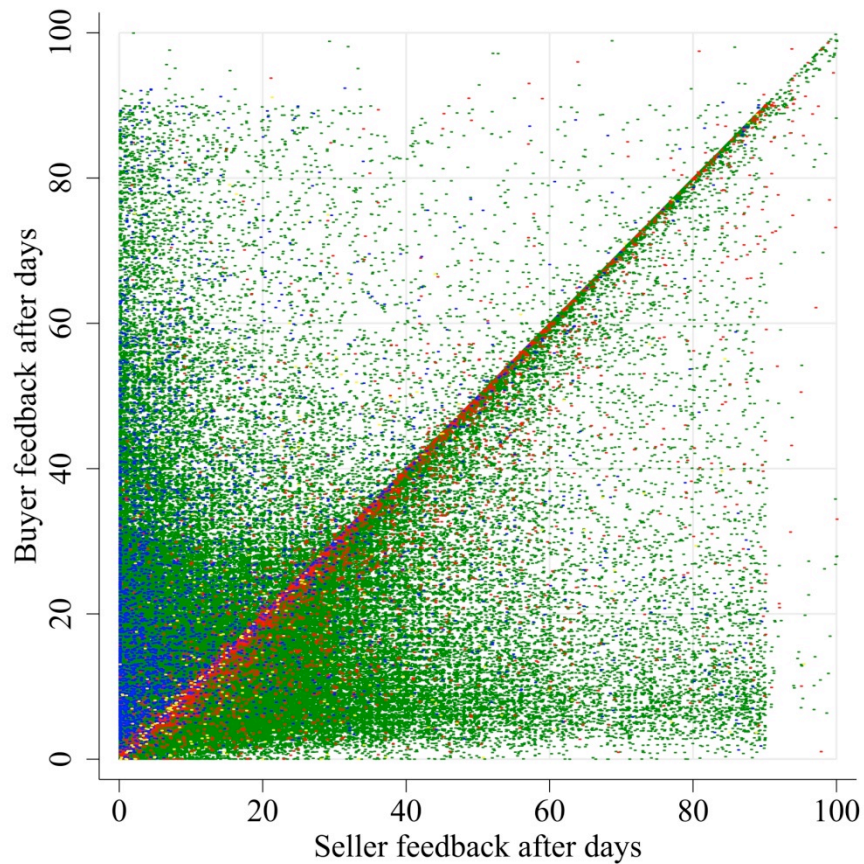
of their feedback period (both for the time before and after the institutional change discussed here), data sets from various eBay platforms in four continents differing in their feedback institutions, as well as data sets from reputation systems of other Internet market platforms, and from markets implemented in the laboratory – as further discussed below.

¹² See <http://ben.orsee.org/supplements> for a description of the field datasets used in this paper. In our description of the field data that motivate our experiment, here as well as in Section V, we report mostly descriptive and simple correlations rather than more in-depth regression analysis. We believe that, given the number of observations and the economic size of the reported effects, such ‘eye-ball tests’ combined with the cited evidence from other studies will be sufficient to convince readers that reciprocity is an issue. Moreover, our laboratory study provides complementary and highly controlled evidence for these phenomena. While not reported here, regressions of feedback behavior (e.g., feedback probability, timing, and content) on observables, controlling for various factors such as country and product category, do confirm our findings (see Ariely et al., 2005, and Kagel and Roth, 2005, for a similar approach of complementing field with laboratory data).

game”). Indeed, the evidence suggests that this kind of reciprocal behavior is an effective tactic for reputation building.¹³

Feedback content. Also observe from Table 1 that there is a high positive correlation between buyer and seller feedback within all countries sampled. There are probably a number of reasons for this; for example, a problematic transaction might leave both sides dissatisfied. But Table 1 also provides a first hint that reciprocity in feedback content has a strategic element: If feedback is given independently, the correlation between seller and buyer content, as measured by tau, should be about the same when the seller gives second as when the seller gives first. In fact, the correlation is about twice as high when the seller gives second. The pattern is similar across countries.

FIGURE 1: CONTENT AND TIMING OF MUTUAL FEEDBACK ON EBAY



Notes: The scatter plot reports about 460,000 observations where both transaction partners gave feedback. ‘Problematic’ feedback includes neutral or withdrawn feedback.

¹³ Based on a model that exploits information about the timing of feedback provision when the partner does not provide feedback, Jian et al. (2010) estimate that buyers and sellers on eBay employed a conditional strategy of giving feedback if and only if the partner did about 20-23% of the time. See also Ockenfels and Resnick (2010) for a selected survey and discussion of this and related literature.

Feedback timing. If feedback timing were independent among trading partners, one would expect the timing of buyer and seller feedback to be uncorrelated with content. But this is not the case: Figure 1 shows the distribution of feedback timing for those transactions where both traders actually left feedback. The green dots represent the timing of mutually positive feedback. More than 70% of all these observations are located below the 45 degree line, indicating that in most cases the seller gives feedback after the buyer. The red dots visualize observations of mutually problematic feedback. Here, the sellers' feedback is given second in even more than 85% of the cases. Moreover, mutually reciprocal feedback is much more heavily clustered alongside the 45 degree line as compared to the case of non-reciprocal feedback. For instance, a seller who gives negative feedback does so much faster after the buyer gave negative feedback than after the buyer gave positive feedback: the median number of days since the buyer gave negative feedback (standard deviation) is 0.77 (11.1), and 2.98 (17.9) if he gave positive feedback. All differences are highly significant at any conventional level. The tightness and sequence in timing suggests that sellers quickly reciprocate positive and 'retaliate' negative feedback (which will be confirmed in our controlled laboratory environment).

Seller retaliation also explains why more than 70% of cases in which the buyer gives problematic feedback and the seller gives positive feedback (blue dots in Figure 1), involve the buyer giving second; not doing so would involve a high risk of being retaliated. Observations in which only the seller gives problematic feedback (yellow dots) are rare and have their mass below the 45-degree line.

Why do sellers retaliate negative feedback? Existing theory and laboratory studies on reputation building, while not developed in the context of the production of reputation information, suggest that there are multiple strategic and social motives at work¹⁴ (and these dovetail well with anecdotal and survey evidence that we have collected). Some retaliation is probably driven by social preferences or emotional arousal, e.g., when a buyer's negative feedback is deemed undeserved by the seller. Retaliating negative feedback may also help to deter negative feedback in the future, because retaliation is viewable by buyers in a seller's feedback history. Also, giving a negative feedback increases the probability that the opponent will agree to mutually withdraw the feedback. In fact, of the cases where a seller responds to a negative with a negative feedback, about 27% are later withdrawn. Yet if the seller gives negative feedback before the buyer's negative is posted (and so does not retaliate), the percentage of mutual withdrawal is only about 3%, suggesting that such negative feedback is driven by different motives.

¹⁴ See, e.g., Kreps and Wilson (1982), Milgrom, North, and Weingast (1990), Greif (1989), Camerer and Weigelt (1988), Neral and Ochs (1992), Brandts and Figueras (2003), and Bolton et al. (2004) for the strategic role in reciprocity, and Fehr and Gächter (2000), as well as the surveys in Cooper and Kagel (forthcoming) and Camerer (2003) for the social aspect in reciprocity. Herrmann, Thöni, and Gächter (2008) provide cross-cultural evidence for anti-social reciprocity in laboratory experiments where high contributors to public goods are punished by low contributors.

II. 3 *Two alternative redesign proposals*

Any institutional change in a running market must respect certain path dependencies. This is particularly true for reputation systems, which by their nature connect the past with the future. For this reason, the redesign proposals we consider maintain (in some form) the conventional ratings of the existing system, allowing traders to basically maintain their reputation built before the change.¹⁵ At the same time, each proposal attacks one or the other of two features that appear to facilitate reciprocal behavior, either the open, sequential posting that allows a trading partner to react to the feedback information, or the two-way nature of the ratings that allows sellers to retaliate buyers.

Proposal 1. Make conventional feedback blind. To do so, conventional feedback would be given simultaneously in the sense that traders cannot see the opponent's feedback before leaving one's own feedback. This way, traders cannot condition their feedback on the feedback posted by a transaction partner, thereby excluding sequential reciprocity and strategic timing, making seller retaliation more difficult. The conjecture is that this will lead to more accurate feedback.¹⁶ A major risk with a blind system concerns whether it will diminish the frequency of feedback giving, particularly with regard to mutually satisfactory transactions. Because trading partners effectively give feedback simultaneously, giving a positive feedback could not be used to induce a trading partner to do the same.

Proposal 2. Add a detailed seller rating (DSR) system to supplement conventional feedback. In principle, a one-sided system in which only the buyer gives feedback is the surest way to end seller retaliation.¹⁷ But while there is more scope for moral hazard on the seller side than on the buyer side in eBay's marketplace, there is some room for buyer moral hazard as well.¹⁸ Moreover, gaining positive feedback as a buyer appears to be an important step for many traders in their transition to a successful seller. For these reasons, the proposal was to create a detailed seller rating system to supplement the conventional feedback system: Conventional feedback would be published immediately, as usual, but the buyer, and only the buyer, can leave additional

¹⁵ Another example for the consideration of path dependency in practical reputation system design can be found on Amazon.com. When changing its ranking of voluntary book reviewers in 2008, Amazon retained its classical system (tracking lifetime quantity of reviews) while adding new measures to reflect the quality of reviews.

¹⁶ A blind system of this sort has been suggested by Güth, Mengel, and Ockenfels (2007), Reichling (2004) and Klein et al. (2007), among others. Miller, Resnick, and Zeckhauser (2005) propose a scoring system which makes reporting honest feedback, in the absence of other feedback-distorting incentives, part of a strict Nash equilibrium, but do not consider the problem of reciprocally biased feedback.

¹⁷ A system of this sort has been proposed by Chwelos and Dhar (2007), among others. These systems share elements with a system that strictly separates feedback earned as a seller and feedback earned as a buyer, which is discussed and experimentally analyzed in Bolton et al. (2010).

¹⁸ While there is little formal evidence at this point for buyer moral hazard on eBay, we collected considerable evidence in our surveys with eBay traders conducted jointly with eBay, anecdotal evidence from eBay's online feedback forum and from eBay seller conferences. There are basically four themes: 1) The buyer purchases the item, but never sends the payment, which incurs time costs on the seller as well as fees to eBay. 2) The buyer has unsubstantiated complains about the item. 3) The buyer blackmails the seller regarding feedback. Stories exist where the buyer asked for a second item for free, threatening negative feedback if this wish would not be fulfilled. 4) After two months the buyer asks the credit card provider to retrieve the payment (eBay's payment service PayPal does not provide support in these cases).

feedback on the seller under blind conditions so that the seller cannot reciprocate them.¹⁹ A major risk is that the conventional and DSR feedback given to sellers might diverge. Unhappy buyers might give positive conventional feedback to avoid seller retaliation, and then be truthful with the (blind) DSR score. This could cause the feedback system scoring to have a new kind of credibility problem with traders.

To summarize this section, people tend to reciprocate the feedback they are given, like with like. The timing and posting rules governing feedback giving in eBay’s conventional feedback system facilitate reciprocity and strategic responses to reciprocity. This has both negative and positive consequences for the system. On the negative side, reciprocity distorts the production and content of feedback information in individual interactions. On the aggregate level, these prospects of positive reciprocal and negative retaliatory feedback may lead to ‘overly’ positive feedback, hampering the informativeness of the system. The fact that from all 742,829 eBay users in Dataset 1, who received at least one feedback, 67% have a percentage positive of 100%, and 80.5% have a percentage positive of greater than 99%, provides suggestive support for the bias. The observation is in line with other eBay research suggesting that feedback is ‘overly’ positive,²⁰ and also with a general tendency for lenient and compressed performance ratings, as discussed for instance in the literature on the “leniency bias” and “centrality bias” in human resource management (Bretz, Milkovich, and Read, 1992; Prendergast and Topel, 1993; Prendergast, 1999). On the positive side, reciprocity appears to be an important motivator in getting mutually satisfactory trades reported. These patterns of positive and negative reciprocity are replicated in our controlled laboratory study (Section IV). The blind conventional feedback proposal alters the sequential, open feedback rule, while the DSR proposal creates an asymmetry in the system so that sellers do not have the opportunity to retaliate the more detailed rating. Each has risks of generating adverse side effects concerning, respectively, the amount or consistency of feedback given.

III. Descriptive evidence from other Internet markets

As a first step in evaluating the two proposals, we searched for and examined systems involving blind and one-sided feedback in other Internet markets. The benefit of field data is that we can study behavior in naturally evolved environments, with real traders. The need for an experiment arises because there are typically other differences than the feedback system (see Kagel and Roth, 2000, for an analogous rationale in

¹⁹ Another advantage is that we can fine tune the scaling of the new ratings without disrupting the 3-point conventional ratings; the latter would create a number of path dependency problems. Research in psychology suggests that Likert scaling of 5 or 7 points is optimal (e.g., Nunnally, 1978; and more recently Muniz, Garcia-Cueto, and Lozano, 2005). Additionally, several studies have found that users generally prefer to rate on more categories rather than submitting just one general rating (e.g., Oppenheim, 2000). The specific method for posting detailed seller ratings is best understood in the context of a number of practical considerations and is described at the beginning of Section V.

²⁰ Dellarocas and Wood (2008) examine the information hidden in the cases where feedback is not given. They estimate that buyers are at least mildly dissatisfied in about 21% of all eBay transactions, far higher than the levels suggested by the reported feedback. Similarly, in a controlled field experiment conducted on eBay with experienced eBay traders, Bolton and Ockenfels (2008) found that sellers, who did not share the gains from trade in a fair manner, received significantly less feedback than sellers who offered buyers a fair outcome.

the context of designing matching market rules for new physicians; our experiment is described in Section IV).

The first field evidence comes from eBay’s own market in Brazil. *MercadoLivre* began in 1999 as an independent market, eBay-like in its objective but with some unique trading procedures. EBay bought the market in 2001 and decided to keep some procedures, including a blind feedback system.²¹ MercadoLivre reveals submitted feedback after a 21-day “*blind period*” that starts upon completion of the transaction. No feedback can be given after the blind period has lapsed. Table 3 shows feedback statistics based on a total of 24,435 completed transactions in Dataset 3, which was specifically compiled to compare feedback behavior in eBay’s conventional feedback system to other eBay sites like MercadoLivre and eBay China, which we will come back to in the next section.²² Observe that the share of problematic (negative, neutral, and withdrawn) feedback given on MercadoLivre is multiple times higher than on other mature eBay platforms that do not employ a blind feedback system, strongly suggesting that blindness indeed affects the feedback distribution. Moreover, while the correlation of feedback content differs little from that in other markets (Column 7 in Table 3), the correlation of feedback giving is much lower in Brazil than in the U.S., Germany, or China (Column 8 in Table 3). That is, in those cases where both transaction partners leave feedback, the content in Brazil is as correlated as in the other countries, but the probability of two-way feedback giving is much smaller.

One worry with a blind system is that diminishing reciprocal opportunities might diminish the rate at which traders leave feedback. But on MercadoLivre there is no evidence that the blind system decreases participation; the feedback frequency of 71% for buyers is in line with what we observe in other countries, and with 88% sellers provide even more feedback.

TABLE 2: FEEDBACK FREQUENCY, CONTENT AND CORRELATION ON MERCADO LIVRE AND EBAY CHINA COMPARED TO OTHER EBAY PLATFORMS

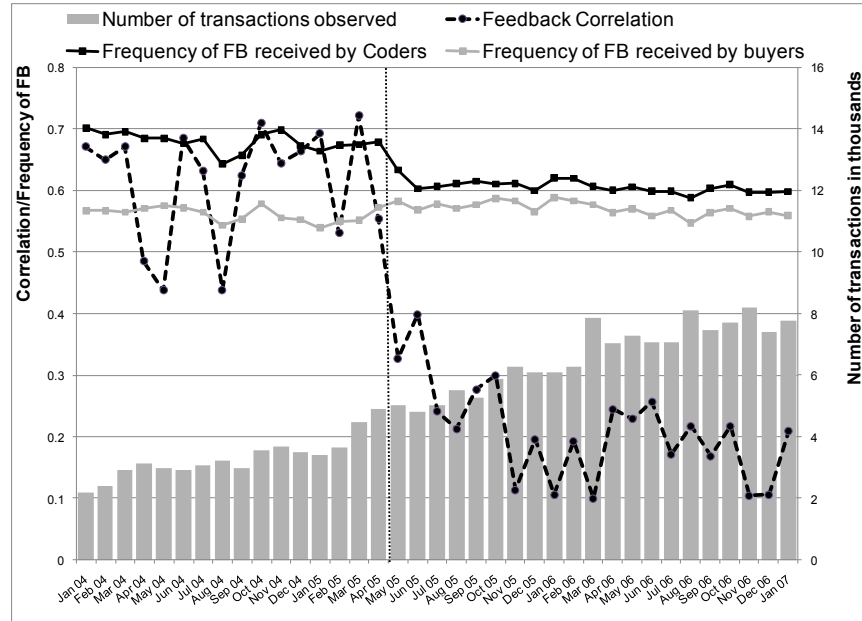
	N	Feedback frequency		problematic FB given by		FB Content Correlation	FB Giving Correlation
		Buyer	Seller	Buyer	Seller	Kendall's tau	Kendall's tau
eBay U.S.	10,169	74.8%	76.7%	1.4%	1.2%	0.720	0.595
eBay Germany	14,297	77.3%	76.9%	1.9%	1.1%	0.621	0.623
eBay China	2,011	9.3%	19.7%	5.0%	6.7%	0.576	0.652
... verified buyers	1,062	15.0%	13.6%	5.0%	4.9%	0.576	0.682
... unverified buyers	949	3.1%	3.6%		14.7%		0.460
MercadoLivre Brazil	1,958	71.1%	87.9%	18.7%	29.2%	0.785	0.175

Note: All correlations are highly significant.

²¹ Hortaçsu, Martínez-Jerez, and Douglas (2009) compare *bidding* behavior on different MercadoLivre sites and eBay.

²² See Footnote 12 for a reference to a detailed description of our datasets. MercadoLivre posts only the day of feedback provision, but not the time, and also updates this date stamp when further verbal comments are left. For this reason we cannot provide information on whether buyer or seller left feedback first.

FIGURE 2: FEEDBACK FREQUENCY AND CORRELATIONS BEFORE AND AFTER THE SYSTEM CHANGE IN APRIL 2005 ON RENTACODER.COM



Overall, the data seem to suggest that the blind system generates a more accurate, or at least a more dispersed reflection of trader satisfaction on both sides of the market. However, as with every comparison across markets, there are a number of potentially confounding effects complicating the comparison. We will discuss this issue at the end of the section, when we have presented the data from other field comparisons.

The *RentACoder.com* site enables software coders to bid for contracts offered by software buyers. RentACoder.com used to have a two-sided, open feedback system, similar to eBay, but switched to a blind system in May 2005. RentACoder’s motive for the switch (as stated on its help page) is the potential threat of retaliatory feedback in an open system. The blind system allows buyers and coders to leave feedback on one another within a period of two weeks after completion of a project.

The RentACoder.com panel data (Dataset 4) comprises 192,392 transactions. Unlike the MercadoLivre comparison, it allows for a within-site comparison, keeping all institutions but the feedback system fixed, and allowing an analysis of the transition from an open to a blind system. The transition has no significant effect on average feedback content received by either buyers or sellers, although there is a weakly significant, small increase in the standard deviation of feedback received by buyers.²³ There are, however, other effects indicative of diminishing reciprocity. First, as shown in Figure 2, the monthly correlation between feedback content sharply and significantly drops from an average of 0.62 in the 15 months before the change to 0.21 in the 21 months after the change. We also observe from Figure 2 (and backed by time series regressions controlling for trends), that coders get significantly less feedback after introduction of blind feedback, while

²³ Because of space limitations, we omit here the regressions of time series of monthly averages on constant, time trend and blindness dummy, which confirm the observation.

buyers get a small but significant increase. Overall, the patterns of reduced interdependencies of feedback giving and content seem to suggest that feedback might be more informative. At the same time, however, there is some evidence from RentACoder.com that feedback frequency may be negatively affected by a blind system.

The hypothesis that reducing reciprocity diminishes feedback giving finds further support by looking at systems with *one-way* feedback characteristics. The first evidence comes from a kind of within-platform comparison on the *Chinese eBay site*, where there is a large proportion of so-called “unverified buyers” – buyers who did not provide proof of their identity (yet). Feedback given by unverified buyers does not count towards the seller’s reputation. From a reciprocity perspective, giving feedback to unverified buyers is not unlike giving one-sided feedback. Table 3 shows frequency and content of feedback for verified and unverified buyers. We observe that verified buyers receive and give about five times as much feedback as unverified buyers ($\chi^2=82.6, p < 0.001$) and that feedback giving is much more correlated with verified buyers (the correlation coefficients are 0.460 versus 0.682).²⁴

More evidence comes from *Amazon.de*, which has a one-sided buyer-to-seller feedback system (Dataset 5, a sample of 320,609 feedbacks).²⁵ In addition, we conducted a small email-based survey with a subset of sellers in our sample. Taking the survey responses of 91 Amazon sellers and the field data together, we find that feedback is left by buyers in about 41% of transactions; if we weight the answers by number of transactions, we get a 36% figure (implying that very active sellers get somewhat less feedback), about half the rate of feedback on the various eBay platforms.²⁶ So, overall, the evidence suggests that the scope for reciprocity affects the decision to give feedback.

While the analysis of the field data sets is suggestive of how the proposed solutions may affect reciprocal feedback and feedback giving, they also raise many questions. Given the highly complex environments these markets operate in, it is difficult to make clear causal inferences. For instance, the low level of positive feedback in MercadoLivre may stem from uncontrolled cross-country effects regarding different norms of trading or feedback giving, or from differently developed payment or postal services. Similarly, a comparison of RentACoder.com with eBay is complicated by the fact that the RentACoder.com feedback is on a 10-point

²⁴ Moreover, unverified buyers receive a neutral or negative feedback with probability 14.7% in our sample, whereas verified buyers receive a negative feedback only with probability 4.9% (due to low feedback frequency, $\chi^2=2.82, p = 0.093$), suggesting that a one-sided system will elicit less positive (and probably more accurate feedback) feedback. However, here, the causality appears to be less clear. Unverified buyers might be more likely to be not familiar with the trading and communication norms on the site or to have less long-term interests on the site and so have less incentive to build up a good reputation.

²⁵ Strictly speaking, both sellers and buyers on Amazon are able to submit feedback on each other. However, feedback given to buyers is not accessible to other sellers, while feedback to sellers is published publicly. As a result, sellers typically do not leave feedback. This makes Amazon’s system effectively a one-sided one.

²⁶ We also observe that Amazon feedback is more discriminative than eBay conventional feedback in the sense that only 81.5% of feedback is given in the best category of 5, while middle and low feedback of 4, 3, 2, and 1 is given in 14.5%, 2.2%, 1.0%, and 0.9% of all cases, respectively.

scale, the market is smaller, the bidding process and price mechanism are different (coders bid for contracts and buyers do not need to select the lowest price offer), and the networking and trading patterns are likely to differ (e.g., there is likely to be little overlap between program buyers and programmers).

Along the same lines, the field data provide no direct evidence that the reduction in reciprocity improves either the informativeness of feedback or market efficiency. One reason to wonder is that the market in the sample closest to the eBay markets in question, MercadoLivre, exhibits a far higher rate of negative feedback than any other market.²⁷ Another reason is the relatively low rates of feedback giving in some of the markets with blind or one-sided feedback: a substantial drop in feedback giving might raise its own credibility issues, effectively substituting one trust problem for another. With the exception of RentACoder.com, there is little in the way of before and after data to guide such an analysis.

Finally, the one-sided proposal, which combines eBay's conventional feedback with detailed seller ratings, has no precedent in Internet markets. One of the major arguments for this proposal is that it represents a more modest shift from the existing system than does the blind proposal or truly one-sided systems such as Amazon's. At the same time, it runs its own unique risk in that, under the system, conventional feedback might diverge from the new, more detailed feedback. Again, this might lead to a loss in credibility.

IV. The laboratory study

The issues discussed at the end of Section III stem from problems in evaluating the proposed redesigns on field data alone; problems that we address with an experiment. Specifically, the experiment is designed as a level playing field for comparing the performance of the different feedback system designs, holding the market environment constant. By the same token, the various experimental controls help us isolating the role of reciprocity for feedback giving and establishing causal relationships between feedback institutions and market behavior. Finally, the experiment provides direct measures for feedback informativeness and market efficiency.

Subsection IV.1 outlines the experimental design. Subsection IV.2 shows that the laboratory feedback behavior we observe mirrors key field observations from the conventional system, and that different systems lead to different feedback behaviors. Subsection IV.3 measures the impact of the feedback system on the economic performance of the auction market. Subsection IV.4 shows how market performance is connected to feedback informativeness.

²⁷ One response to this concern is that the rate of negative feedback on MercadoLivre accords well with rates of unhappiness uncovered by research (ex., Dellarocas and Wood, 2008). However, as the experiment reported in the next section makes clear, we should expect more informative feedback to ignite a number of endogenous effects in the system, starting with buyers better indentifying and shunning untrustworthy sellers, and so the proportion of unsatisfactory trades should be something less than the present rate of unhappiness.

IV.1 Experimental design and a hypothesis

The experiment simulates a market with an auction component (including a moral hazard element) that was held fixed across all treatments, and a feedback component that was varied to capture the various scopes for reciprocity across the alternative systems.²⁸

Auction component. Each treatment simulates a market that consists of 60 rounds. In each round participants are matched in groups of four, one seller and three potential buyers. Each buyer i receives a private valuation for the good, v_i , publicly known to be independently drawn from a uniform distribution of integers between 100 and 300 ECU (Experimental Currency Units). Buyers simultaneously submit bids of at least 100 ECU or withdraw from bidding. The bidder with the highest bid (earliest bid in case of a tie) wins the auction and pays a price p equal to the second highest bid plus a 1 ECU increment, or his own bid, whichever is smaller. If there is only one bid, the price is set to the 100 ECU start price. After the auction, all participants in the group are informed of the price and of all bids but the highest.²⁹ The price is shown to the seller s who then determines the quality of the good $q_s \in \{0, .01, \dots, .99, 1\}$.³⁰ The payoff (not including feedback costs described below) to the seller is $\pi_s = p - 100q_s$ and to the winning buyer i is $\pi_i = q_i v_i - p$.

There were 32 participants in a session and two sessions per treatment. Eight sequences of random parameters (valuations, role and group matching), involving 8 participants each, were created in advance. Thus, random group re-matching was restricted to pools of 8 subjects, yielding four “sub-sessions” per session and 8 statistically independent observations per treatment. To ensure a steady growth of experience and feedback, random role matching was additionally restricted such that each participant became a seller twice every 8 rounds. The same 8 random game sequences were used in all treatments. Participants were not informed about the matching restriction.

Feedback component. When the auction ends in a trade, both buyer and seller have the opportunity to give voluntary feedback on the transaction partner. Giving feedback costs the giver 1 ECU, reflecting the small effort costs when submitting feedback.

In the *Baseline* treatment, both the seller and the buyer can submit conventional feedback (CF), rating the transaction negative, neutral, or positive. Feedback giving ends with a “soft close”: In a first stage, both transaction partners have the opportunity to give feedback. If both or neither give feedback, then both are informed about the outcome and the feedback stage ends. If only one gives feedback, the other is informed about that feedback and enters the second feedback stage where he has again the option to give feedback, and

²⁸ See Appendix A.1 for a translation of experimental instructions.

²⁹ For simplicity, we chose a sealed-bid format and abstracted away from eBay’s bidding dynamics, which is known to create incentives for strategic timing in bidding (Roth and Ockenfels, 2002). Other features such as the handling of increments and the information feedback are chosen analogously to eBay’s rules.

³⁰ In the experiment, sellers were asked to choose an integer between 0% and 100%. Proportions simplify the notation. The quality choice used in our experiments is a simplification of the many potential dimensions of moral hazard in the field, like inaccurate item descriptions, long delivery time, low quality, etc.

so a chance to react to the other's feedback.³¹ As on eBay, a trader's conventional feedback is aggregated over both buyer and seller roles as the feedback score and the percentage of positive feedbacks (cf. Section II). When the participant becomes a seller, these scores are presented to potential buyers on the auction screen prior to bidding.

The *Blind* treatment differs from the *Baseline* only in that we omit the second feedback stage. That is, buyer and seller give feedback simultaneously, not knowing the other's choice.

The *DSR* (Detailed Seller Rating) treatment adds a rating to the *Baseline* treatment feedback system. After giving CF, the buyer (and only the buyer) is asked to rate the statement "The quality was satisfactory" on a 5-point Likert scale: "I don't agree at all", "I don't agree", "I am undecided", "I agree", "I agree completely". As in the *Baseline* treatment, we implement a soft close design, but in case the seller delays and enters the second feedback stage, she is only informed about the conventional feedback given by the buyer, not about the detailed quality rating. Number and average of received detailed seller ratings are displayed at the auction page.

All sessions took place in April 2007 in the Cologne Laboratory for Economic Research. Participants were recruited using an Online Recruitment System (Greiner, 2004). Overall 192 students participated in 6 sessions.³² After reading instructions and asking questions, participants took part in two non-interactive practice rounds (see Appendix A). Each participant received a starting balance of 1,000 ECU to cover potential losses. Sessions lasted between 1½ and 2 hours. At the end of the experiment, the ECU balance was converted to Euros at a rate of 200 ECU = 1 Euro, and was paid out in cash. Participants earned 17.55 Euros on average (standard deviation = 2.84), including a show-up fee of 2.50 Euros and 4 Euros bonus for filling in a post-experiment questionnaire.

The experiment is designed to isolate the reciprocal relationship between traders. In particular, the experiment abstracts away from any buyer moral hazard: each winning bid is automatically transferred to the seller. Thus, the feedback given by sellers to buyers is not informative, because it cannot represent reputational information on the trader's behavior as a seller.³³ Yet, as on eBay, leaving feedback on a buyer may affect the buyer's future profits as a seller, because a trader's reputation depends on feedback earned both as a buyer and a seller. So positive feedback may be effectively rewarded and negative feedback may be retaliated.

³¹ Klein et.al. (2007), as well as our data, support this characterization of the conventional eBay system. Ariely, Ockenfels, and Roth (2005) and Ockenfels and Roth (2006) model the ending rule of Amazon.com auctions in a similar way, allowing buyers to always respond to other bids.

³² According to their answers to a post-experimental questionnaire, the average age of our participants was 23.8 years, 49% of them were male, two-thirds were eBay members. On average each had bought (sold) 35.3 (23.5) items on eBay.

³³ Insofar buying behavior is uncorrelated with selling behavior, this also holds for eBay.

Hypothesis. The experiment has a finite number of trading rounds. Assuming that all agents are commonly known to be selfish and rational, the unique subgame-perfect equilibrium in all treatments of the experiment stipulates zero feedback giving and quality tendered, with no auction bids. The socially efficient outcome has the bidder with the highest valuation winning the auction, the seller producing 100% quality, with no (costly) feedback giving. So both of these, rather extreme, scenarios leave no role for the feedback system. If, as seems more likely, feedback is used to build up reputation and to discriminate between sellers, we hypothesize that reciprocal feedback hampers market efficiency because reciprocity compresses reputation scores in a way that makes it harder for buyers to discriminate between sellers; these sellers then have less incentive to deliver good quality. While the contribution of this paper is not theory (although our empirical findings may prepare the ground for new theory; see Section VI), Appendix B illustrates in the context of our experimental setup how a bigger distortion induced by reciprocity might lower shipped quality, bids, prices, and market efficiency.³⁴ Consequently, the two proposed redesigns, if they diminish the role of reciprocity, should do better.

TABLE 3: TIMING OF FEEDBACK

	<i>Baseline</i>	<i>Blind</i>	<i>DSR</i>
Both first round	27%	26%	29%
None first round	16%	24%	15%
Seller 1 st , buyer in 2 nd	4%		2%
Seller 1 st , Buyer not (in 2 nd)	5%	8%	8%
Buyer 1 st , seller in 2 nd	24%		17%
Buyer 1 st , Seller not (in 2 nd)	23%	42%	28%

TABLE 4: KENDALL TAU CORRELATIONS
BETWEEN SELLER AND BUYER FEEDBACK BY TIMING

	Both 1 st	S 1 st , B 2 nd	B 1 st , S 2 nd	All
<i>Baseline</i>	0.359	0.536	0.901	0.680
<i>Blind</i>				0.411
<i>DSR</i>	0.533	0.730†	0.913	0.759

Note: All correlations highly significant at the 0.1% level, except for cell indicated by † which is weakly significant at the 10% level.

³⁴ We do this with the help of a simple example, assuming away several behavioral and institutional complexities and not explicitly modeling the underlying motivational mechanisms for reciprocity. We concentrate instead on what we believe captures the essence of the effect of distorted feedback on efficiency both in the lab and on eBay. For an overview on different modeling approaches to seller reputation see Bar-Isaac and Tadelis (2008); for earlier related work see Klein and Leffler (1981). There is also an experimental literature testing reputation theory; more recent contributions include Grosskopf and Sarin (2010), allowing for reputation to have either beneficial or harmful effect on the long-run player (partly based on a model by Ely and Valimaki, 2003), and Bolton et al. (2011), searching for information externalities in reputation building in markets with partners and strangers matching, as predicted by sequential equilibrium theory for some games. These as well as other papers (see references cited therein) come to the conclusion that reputation building often interacts with social preferences in subtle ways, often (but not always) making reputation mechanisms more beneficial than predicted by theory based on selfish behavior. Our study complements this literature by showing how reciprocity can both hamper and promote the effectiveness of reputation mechanisms.

IV.2 Feedback Behavior

In this section, we investigate whether the feedback pattern in the *Baseline* treatment mirrors the pattern observable in the field, and how the feedback behavior in the alternative systems compares.

Feedback giving. In the *Baseline* treatment, buyers give feedback in about 80% and sellers in about 60% of the cases, with an average of about 70%. Relative to *Baseline*, *Blind* exhibits significant drops in both buyer (68%) and seller (34%) giving frequencies (two-tailed Wilcoxon $p < 0.025$ in both cases), whereas *DSR* exhibits only minor and insignificant reductions for both buyers (77%) and sellers (57%; $p > 0.640$ in both cases).³⁵

Feedback timing. When possible, sellers are more likely than buyers to wait until the other has given feedback (Table 4; Wilcoxon two-tailed $p < 0.025$ both in *Baseline* and *DSR*). This effect is most pronounced when feedback is mutually neutral/negative; the only case with buyers more often moving second is – like in the field data – when the buyer gives a problematic and the seller a positive conventional feedback (see Table 10 in Appendix C for details). This interaction pattern of feedback content and timing is very similar to what is observed in the field (Section II), and thereby reassures us of the suitability of our experimental design of the *CF* feedback component.

Feedback content. Table 5 shows correlations between conventional feedbacks across treatments. We find that blindness of feedback significantly decreases the correlation compared to the open systems. The high correlations in the latter are mainly driven by the cases where sellers delay their feedback and give second, while when both transaction partners give feedback in the first stage, correlations are comparable to blind feedback. However, correlations of simultaneously submitted feedback are significantly different from zero, too.

Negative feedback. Finally, the probit estimate in Table 6 shows the determinants of problematic feedback given to sellers conditional on the buyer giving feedback (where, as before, problematic feedback is defined as either a negative or neutral feedback). Controlling for quality, price and other factors, we see that problematic conventional feedback increases in both *Blind* and *DSR*. The coefficient estimates for the two treatment dummies are nearly identical, indicating that the size of the effect is about the same in both treatments.³⁶

The reason for more negative feedback is that buyers receiving poor quality are more likely to give problematic feedback under the alternative systems. More specifically, Figure 3 illustrates that in all

³⁵ Regression analyses considering interaction effects of treatments with quality support the finding (Table 9 in Appendix C), and furthermore show that buyers give significantly more often feedback when quality is low in both alternative designs. We discuss feedback giving correlations below.

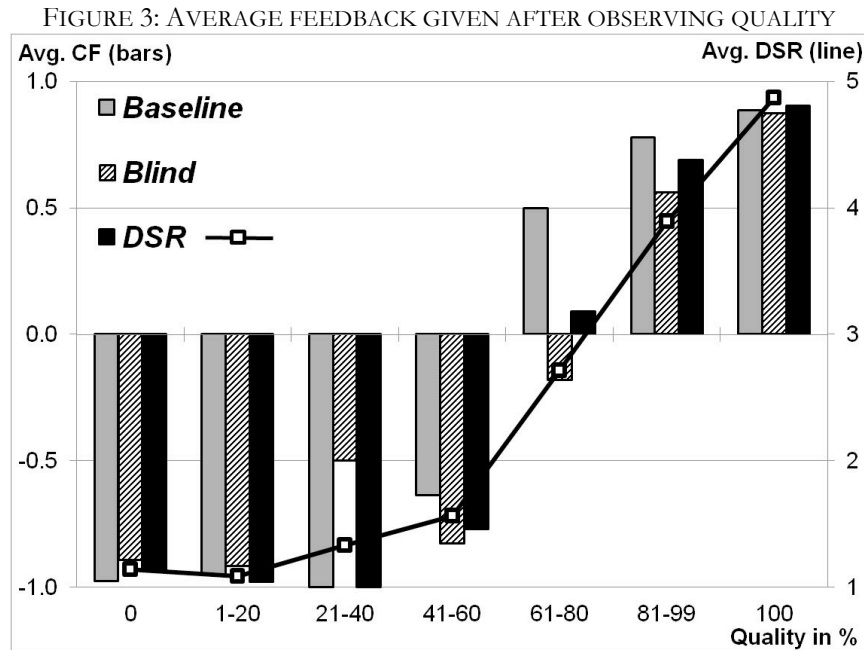
³⁶ The same probit, run on all successful auction data (not conditional on the buyer giving feedback), yields similar results save the coefficient for the *Blind* treatment is somewhat smaller (still positive) but insignificant, most likely because of the drop in feedback frequency we observed earlier for that treatment. The share of positive (negative) buyer-to-seller feedback is 53% (44%) in *CF*, 47% (48%) in *Blind*, and 55% (37%) in *DSR*. Differences in these numbers need to be cautiously interpreted, however, because of simultaneous endogenous changes in the success of sales, the level of quality, and the responsiveness and quality of feedback between treatments, which are controlled for in the regressions and figures in the main text.

treatments, a positive conventional feedback (and the highest DSR) is awarded for quality of 100%; likewise, very low quality receives negative feedback in all cases. The major difference between the treatments happens between 40% and 99% quality; here average conventional feedback given is tougher in *Blind* and *DSR*. In this range, conventional feedback given in *DSR* is not quite as tough, but observe that the DSRs given generally line up well with the *Blind* conventional feedback; that is, the DSRs reflect buyer standards similar to those revealed in *Blind*.

TABLE 5: DETERMINANTS OF PROBLEMATIC FEEDBACK, PROBIT COEFFICIENT ESTIMATES
(ROBUST STANDARD ERRORS CLUSTERED ON MATCHING GROUP, ROUNDS 1 TO 50)

Dep var	Buyer gives problematic feedback	
	Coeff	(StdErr)
Constant	2.770 ***	(0.538)
<i>Blind</i>	0.414 **	(0.204)
<i>DSR</i>	0.417 **	(0.197)
Round	-0.010 **	(0.004)
Price	0.004 ***	(0.001)
Quality	-0.047 ***	(0.008)
S Feedback Score	-0.031	(0.021)
N	1725	
Restricted LL	-558.8	

Note: *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively. *Blind* and *DSR* are treatment dummies. S Feedback Score denotes the feedback score of the seller.



Note: For this figure, CF is coded -1 (negative), 0 (neutral), and 1 (positive). DSR is given on a 1-5 integer scale.

Summing up, the *Baseline* treatment qualitatively replicates the pattern of strategic timing, retaliation and correlation of feedback found on eBay.³⁷ Moreover, as predicted, the alternative systems successfully mitigate reciprocity (as shown, for instance, by reduced correlations of feedback content) and so allow for a more negative response to lower quality.

IV.3 Quality, Prices, and Efficiency

The hypothesis underlying our redesign efforts is that the extent to which feedback is shaped by reciprocity affects economic outcomes. More specifically, we hypothesize that diminishing the role of reciprocity increases quality, prices and efficiency (see Appendix B). Figure 4 shows the evolution of quality and auction prices over time: both quality and prices are higher in both *DSR* and *Blind* than in *Baseline*. Applying a one-tailed Wilcoxon test using independent matching group averages, the increases in average quality and price over all rounds are significant for treatment *DSR* ($p = 0.035$ and 0.025 , respectively), but not for *Blind*. The test, however, aggregates over all rounds, and there is a sharp end game effect in all treatments, with both quality and prices falling towards zero, consistent with related studies on reputation building in markets.³⁸ Regressions controlling for round and end game effects yield strong and significant positive treatment effects regarding bids and quality for both *DSR* and *Blind*, with similar magnitude (see Price Model 1 and Quality Model 1 in Table 7).³⁹

The choice of bid and quality levels affects efficiency. In the *Baseline* treatment, 47% of the potential value was realized, with losses of 23% and 31% resulting from misallocation and low quality, respectively.⁴⁰ Both alternative systems increase efficiency, yet only *DSR* does significantly so; there is a 27% increase in efficiency in *DSR* ($p = 0.027$) compared to *Baseline*, and a 16% increase in *Blind* ($p = 0.320$).⁴¹ Both market sides gain (although not significantly so) in the new system: about 45% (56%) of the efficiency gains end up in the

³⁷ There are only two major exceptions. First, there is no endgame effect in the field, because there is no endgame. Second, we have much more negative feedback in all our treatments compared to eBay. This is wanted, because given the rareness of negatives on eBay it would otherwise be difficult to study reciprocal feedback giving and especially feedback retaliation. See also the previous footnote.

³⁸ See for example Selten and Stöcker (1986) and Bolton et al. (2004).

³⁹ Models 1 and 2 in Table 7 reveal another reciprocity effect, resembling what is frequently observed in trust games: sellers respond to higher price offers with better quality (a 1 ECU price increase comes with a 0.2 percentage point increase of quality). While there is evidence from a controlled field experiment conducted on eBay suggesting that both eBay buyers and sellers may care about reciprocal fairness (Bolton and Ockenfels 2008), we are not aware of any field eBay study investigating whether the final auction price reciprocally affects seller behavior.

⁴⁰ A misallocation occurs if the bidder with the highest valuation does not win so that welfare is reduced by the difference between the highest valuation and the winner's valuation (which we define as the seller's opportunity cost of 100 when there is no winner because of lack of bids). Low quality leads to an efficiency loss because each percent quality the seller does not deliver reduces welfare gains by one percent of the auction winner's valuation, minus one. Also, each feedback reduces welfare by 1 ECU, but this source of efficiency loss is negligible as in no treatment feedback costs exceed 1% of maximal efficiency.

⁴¹ Reported p -values are from a one-tailed Wilcoxon matched pairs signed ranks test on fully independent group averages.

sellers' pockets in *DSR (Blind)*, and the rest goes to buyers. So both alternative systems increase price, quality and efficiency. *DSR* improvements are economically and statistically significant, while efficiency improvement in *Blind* is not significant.

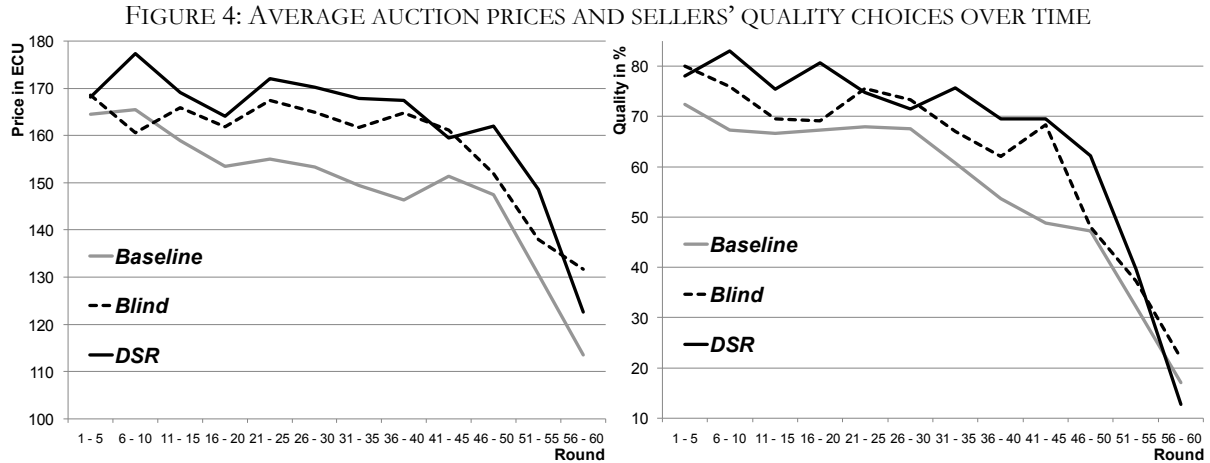


TABLE 6: DETERMINANTS OF QUALITY AND PRICE, TOBIT COEFFICIENT ESTIMATES
(ROBUST STANDARD ERRORS CLUSTERED ON MATCHING GROUP, ROUNDS 1 TO 50)

Dep var	Quality				Price			
	1		2		1		2	
Model	Coeff	(StdErr)	Coeff	(StdErr)	Coeff	(StdErr)	Coeff	(StdErr)
Constant	36.59 ***	(9.659)	45.75 ***	(9.053)	158.85 ***	(5.923)	166.93 ***	(4.795)
<i>Blind</i>	20.75 ***	(7.813)			20.22 **	(8.685)		
<i>DSR</i>	21.27 ***	(5.105)			12.57 **	(6.321)		
Round	-1.17 ***	(0.186)	-1.13 ***	(0.210)	-1.24 ***	(0.144)	-1.18 ***	(0.157)
S FScore			3.85 ***	(0.977)			5.63 ***	(0.944)
S FScore* <i>Blind</i>			3.38 **	(1.696)			3.41 ***	(0.807)
S FScore* <i>DSR</i>			1.05	(1.200)			-0.604	(1.013)
S DSR Avg			6.22 ***	(1.970)			3.75 **	(1.847)
Price	0.216 ***	(0.045)	0.222 ***	(0.045)				
N	2283		2283		2283		2283	
Restricted LL	-7944.2		-7933.2		-11032.8		-11038.7	

Note: *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively. *Blind* and *DSR* are treatment dummies. S FScore denotes the (conventional) feedback score of the seller, and S DSR Avg the seller's average DSR score.

TABLE 7: DETERMINANTS OF SELLER AVERAGE PROFIT, TOBIT COEFFICIENT ESTIMATES
(ROBUST STANDARD ERRORS CLUSTERED ON MATCHING GROUP, ROUNDS 1 TO 50)

Dep var Model	Seller average future profit			
	1		2	
	Coeff	(StdErr)	Coeff	(StdErr)
Constant	73.61 ***	(4.128)	70.45 ***	(4.442)
S FScore			3.04 ***	(0.489)
S FScore* <i>Blind</i>			1.36 *	(0.748)
S FScore* <i>DSR</i>			-2.30 ***	(0.763)
S DSR Avg			3.92 ***	(1.083)
Quality* <i>Baseline</i>	0.079	(0.083)	-0.019	(0.056)
Quality* <i>Blind</i>	0.175 **	(0.082)	0.098	(0.062)
Quality* <i>DSR</i>	0.179 ***	(0.0478)	0.034	(0.042)
Nosale	-53.92 ***	(5.00)	-43.75 ***	(6.533)
N	2400		2400	
Restricted LL	-11398.2		-11180.5	

Note: *, **, and *** indicate significance at the 10%, 5% and 1% level, respectively. *Blind* and *DSR* are treatment dummies. S FScore denotes the feedback score of the seller, and S DSR Avg the average DSR score. The Period variable is omitted because the associated coefficient is small and insignificant.

IV.4 The relationship between feedback informativeness and improvement in market performance

We have seen in Subsection IV.2 that the alternative systems lead to less reciprocal feedback, and in Subsection IV.3 that they lead to improved market outcomes. But how does less reciprocity actually translate into better market performance? The natural hypothesis is that, for a given quality, less reciprocity in feedback giving generates reputation scores that allow better forecasting of sellers' future behavior (see the model in Appendix B). That this is so is evident from Quality Model 2 in Table 7. The model shows that seller conventional feedback scores in *Blind* have significantly higher positive correlation with the quality the seller provides at that point than is the case in *Baseline*. The positive correlation between quality and conventional feedback scores increases in *DSR* as well, but not significantly so. Observe, however, that the DSRs are significantly positively correlated with quality and so, in this sense, the *DSR* seller scores, as well as those in *Blind*, exhibit less distortion than those in *Baseline*.

We expect that introducing one of the alternative systems leads sellers to react to better feedback informativeness by shipping higher quality in *Blind* and *DSR* compared to *Baseline* (see the model in Appendix B). Returning again to Table 7, the Price Model 2 shows that nominally equivalent conventional feedback scores in *Baseline* and *Blind* lead to higher prices in the latter case. In comparing *Baseline* and *DSR*, there is little difference in regard to conventional feedback impact on price; however, DSRs are significantly positively

correlated with price, and in this sense sellers with good feedback scores are more highly rewarded in *DSR* than in *Baseline*.

More evidence comes from looking directly at the effect a quality decision has on a seller’s future average profit. Model 1 in Table 8 shows that the amount of quality a *Baseline* seller chooses in the present round drives up future average profit, but not significantly so. In contrast, the amount of quality a *Blind* or *DSR* seller chooses drives up their future expected profit by a higher and significant amount, and by about the same amount for both treatments.⁴²

To conclude, the experiment mirrors the feedback pattern on eBay pretty well. This gives us confidence that the experiment captures relevant aspects of the field behavior – even though it abstracts away from buyer moral hazard and other features such as mutual feedback withdrawal that may confound the analysis of feedback giving in the field. Data from other market platforms such as RentACoder.com and MercadoLivre suggest that, in particular, blindness might reduce feedback correlation and feedback giving. The experiment replicates these findings in a highly controlled environment, showing that the change of institutions is in fact causal to the observations – although other factors like cross-cultural behavior differences may add to the field patterns. Taken together, the lab and field evidence thus provide a coherent picture of the role of reciprocity in feedback systems. In addition, the experiment complements the field data both by measuring variables that are unobservable in the field and by test-bedding designs that are nonexistent in the field. More specifically, the experiment shows that both systems significantly increase the informativeness of the feedback system, and how the effect is attributable to changes in feedback, bids relative to valuation, and product quality, most of which are difficult to observe in the field. Regarding design, *DSR* yields significant efficiency gains over the baseline treatment and does not experience a significant drop in the number of feedbacks. Because this is not the case for *Blind*, at least not significantly so, the experiment suggests that *DSR* would be the better option for a system change.

V. A first look at the field implementation of detailed seller ratings

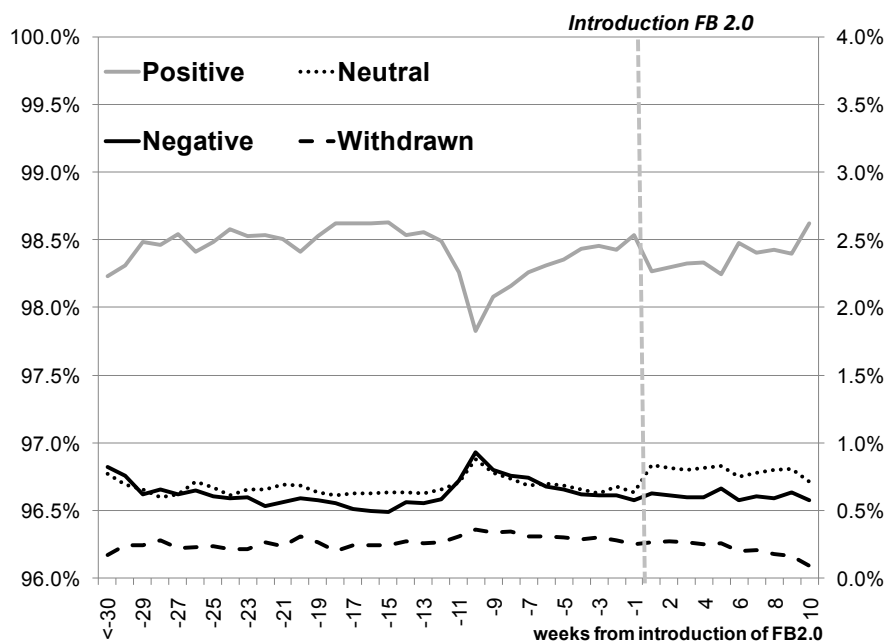
Together, the laboratory and field analyses described above suggest that the *DSR* system outperforms the traditional system and does no worse – and along some dimensions better – than a blind system. Because of this and because of the path dependency concerns mentioned in Section II.3, eBay decided to go for a detailed seller rating feedback system under the name “Feedback 2.0” in spring 2007.⁴³ Under Feedback 2.0, in addition to the conventional feedback, buyers can leave ratings in four dimensions on a 5 point scale.

⁴² As a side note, observe that Model 2 in Table 8 shows that knowing a seller’s feedback score has greater value for forecasting a seller’s future average profit than does knowing the quality decision they make in the present round, in all three treatments. That is, a summary statistic of a seller’s feedback history is a better predictor of his future profitability than observing directly what he did in the present.

⁴³ eBay piloted the new design in a couple of smaller and medium size eBay markets from early March 2007 (Australia, Belgium, France, India, Ireland, Italy, Poland, and UK), and introduced it worldwide in the first week of May 2007.

These dimensions are “How accurate was the item description?”, “How satisfied were you with the seller’s communication?”, “How quickly did the seller ship the item?”, and “How reasonable were the shipping and handling charges?” For each of these ratings, only the number of feedbacks and the average rating are displayed on the seller’s feedback page, and only after the seller receives at least 10 ratings.⁴⁴ On the feedback submission page eBay emphasizes that only averages and no individual DSRs can be observed. As a result, DSR is not only blind (in the sense that it cannot be responded to) and one-sided (only buyers can give detailed ratings), but also anonymous (sellers cannot identify the DSR provider). In this section we present early evidence on the performance of the new system.

FIGURE 5: EVOLUTION OF POSITIVE, NEUTRAL, NEGATIVE AND WITHDRAWN FEEDBACK BEFORE AND AFTER INTRODUCTION OF FEEDBACK 2.0



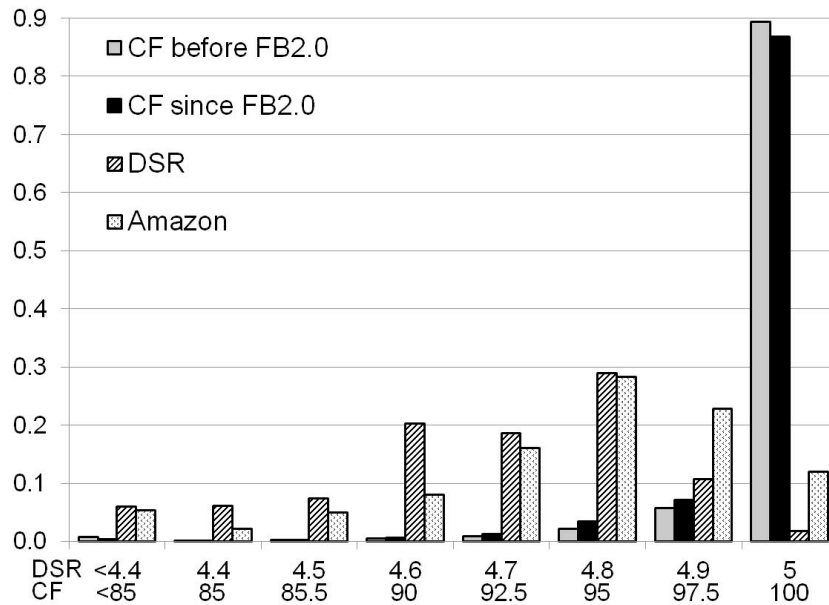
Notes: The figure is based on about 7 and 3 million individual feedbacks in the 30 weeks before and the first 10 weeks after introduction of Feedback 2.0, respectively, in the pilot countries Australia, Belgium, France, Poland and UK. Positive feedback is plotted on the left y-axis, all other feedback on the right y-axis.

The first observation is that conventional feedback giving is not much affected by the system change. Based on our Dataset 1, Figure 6 shows the share of positive (left y-axis) as well as neutral, negative and eventually withdrawn feedback (right y-axis) for the last 30 weeks before and the first 10 weeks after introduction of DSR in early March 2007 (vertical dashed line). The shares are quite stable, with the exception of the kink about 10 weeks before the system change, which falls in the pre-Holiday shopping time known for high expectations and time pressure. From the week before to the week after DSR introduction we

⁴⁴ See Figures 9 and 10 in Appendix C for screen shots. Similar to conventional feedback, DSRs are averaged for each buyer before being aggregated. Also, DSRs older than 12 months are ignored, yielding a ‘rolling’ average. There are a number of other small changes implemented jointly with Feedback 2.0. For instance, information about item title and price were added to feedback comments received as a seller.

observe a small drop in positive and an accompanying rise in neutral feedback. This is in line with the experimental results on the *DSR* system, where we also observe a shift from positive to neutral feedback. However, these changes are small compared to the Holiday shock and overall variance, and seem not to be persistent, at least for positive feedback. We also do not observe significant changes in CF (conventional feedback) frequency, timing or correlation between the pre- and post change Datasets 1 and 2. We conclude that, overall, there are no or at best small short-term effects in CF due to the introduction of DSR.

FIGURE 6: DISTRIBUTION OF AVERAGE CF AND DSR SCORES IN MEMBER PROFILES



Notes: DSR and Amazon.com's 1-5 range and CF percent positive's 0-100 range are divided in the same number of categories and are aligned at the x-axis. eBay data is based on the feedback of the same 27,759 members from Australia, Belgium, France, Poland and UK, received as seller in Jan/Feb 2007 and March/Apr/May 2007, respectively. Inclusion criterion was more than 10 DSRs in at least one DSR category. Amazon data is based on 9,741 Amazon market place sellers.

DSRs are given in about 70% of the cases CF is given, varying somewhat by country and category. Further analysis of conditional DSR frequencies (see Table 11 in Appendix C) shows that DSR feedback is given least often (64%-71%) if the CF feedback is negative, and most often (77%-79%) if the CF is neutral.

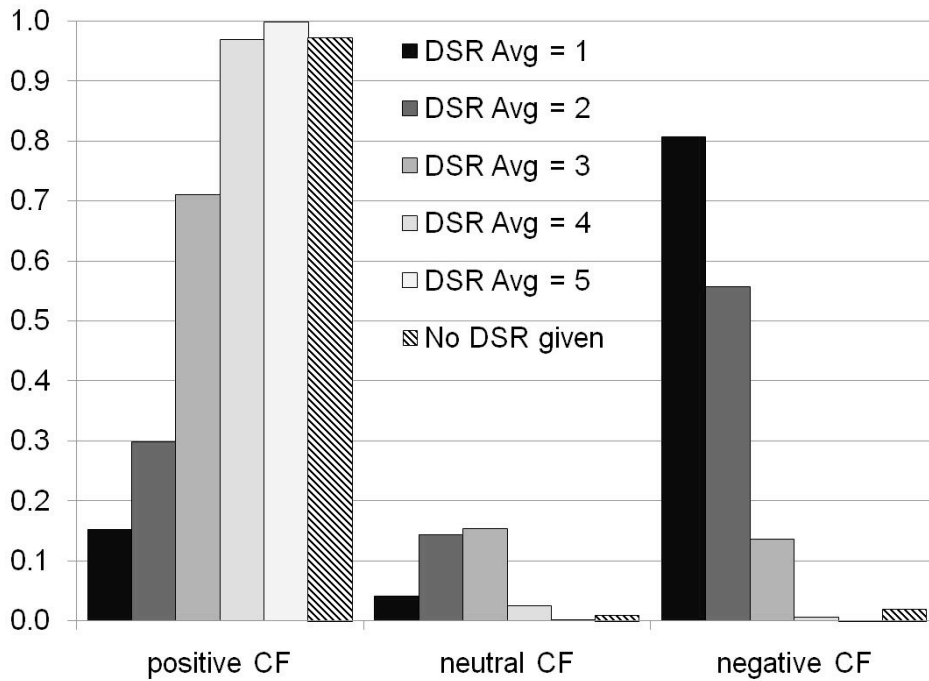
For the 27,759 eBay members from Australia, Belgium, France, Poland and UK in Dataset 1, who received at least ten DSRs between the first week of March and data collection in May 2007 (such that their DSR average was published on their feedback profile), we track CF received as a seller in the same period as well as in the two months before DSR introduction (using individual feedback data from Dataset 1). From this feedback we calculated the fictitious percentage positives of CF of each individual seller before and after introduction of Feedback 2.0, using only those feedbacks given in the corresponding time windows.

In line with Figure 5 above, Figure 6 shows that the CF percentage positives scores slightly decreased after introduction of the new system. However, DSRs are more nuanced. For instance, while most sellers

have a ‘perfect’ CF reputation of 100%, only very few have a ‘perfect’ average DSR of 5. For comparison we also include the distribution of average scores of Amazon.com marketplace sellers in Figure 6 (based on Dataset 5). The one-sided DSR feedback distribution follows the one-sided Amazon.com feedback distribution fairly closely, although it seems to be even somewhat more negative. This supports the idea that DSR is treated as a one-sided system, with little scope for reciprocity. In fact, Figure 7 shows that the difference in rating variability between CF and DSR is partly driven by a strategic response to the differences in the scope for reciprocal behavior.

Figure 7 (based on Dataset 2) shows for each DSR (averaged over the 4 categories) the distribution of the corresponding CFs. As one might expect, when the DSR is 5, virtually all CF is positive, and when the DSR is 4, almost all CF is positive. However, of those buyers who submit the minimum DSR average of 1 (that means that the buyer gave one of the rating combinations 1, 11, 111, 112, 1111, or 1112), about 15% submit a positive CF. For DSR averages of 2, this share is 30%. That is, among those who are maximally unsatisfied measured by DSR, which cannot be reciprocated, a substantial share expresses satisfaction with respect to CF, which can be reciprocated. It seems plausible that at least part of this pattern can be interpreted as hiding bad detailed seller ratings behind a positive open conventional feedback.

FIGURE 7: DISTRIBUTION OF CF CONDITIONAL ON AVERAGE OF CORRESPONDING DSRs



Notes: To calculate the DSR average we take all available of the up to four DSR ratings per feedback, average and round to integer. Thus, a DSR average of 1 implies two or three ratings of 1 and at most one rating of 2.

The initial concern that this kind of strategic hiding behavior might yield inconsistencies between aggregate CFs and DSRs is not borne out, however. The overall share of DSR averages of 1 or 2 is only slightly less than 2%, so that on average, a positive CF comes with a better DSR.

Strategic feedback hiding is only effective when the seller is not able or not willing to retaliate against such feedback. However, while DSR makes retaliation more difficult, one might still suspect that by permanently observing the changes of average ratings, a seller might be able to identify the buyer behind a given DSR. This hypothesis is not supported by our data. When the buyer gives an average DSR of 1 but a positive CF, the probability that the seller retaliates upon this with a negative CF is 0.004, compared to a retaliation probability of 0.468 when the CF is negative.⁴⁵

The experiment suggests that most of the endogenous improvement in performance can be expected from pickier buying. In fact, there is evidence indicating that buyers are indeed more distinguishing under DSR. That is, sellers with a relatively good DSR score have a higher probability of selling listed items after introduction of DSR than the same sellers before introduction of DSR, and sellers with a relatively low DSR have a lower probability of selling with DSR.⁴⁶

VI. Conclusions and challenges for future research

This study is a first exploration of the market design issues surrounding the engineering of trust and trustworthiness in the marketplace. The study illustrates how gaming in the production of reputation information can significantly hamper the ability of a reputation system to facilitate trust and trade efficiency. Our analysis began with the observation that reciprocity plays a major role in the leaving, timing and content of feedback. While retaliatory feedback is in itself a rather small phenomenon, accounting for less than 1.2% of the total mutual feedback data (Figure 1), the *threat* of retaliatory negative feedback distorts feedback in aggregate. The reason is that buyers respond strategically to the threat, either by not reporting bad experiences or waiting for the seller to report first. This, in turn, reduces the informativeness of feedback information, with the end result that a seemingly small phenomenon can substantially hamper trust and market efficiency.

While our paper does not provide new theory, it speaks to the need for new theory. Our method was to observe the phenomenon in the field as best we could, to design a laboratory study to probe the phenomenon in greater detail and to establish causalities, and then to draw analogies based on the robust findings from field and experimental data. These analogies put the phenomenon in sharper relief and suggest data

⁴⁵ In support of this observation, straightforward regression analyses show a very high correlation between seller's and buyer's CF, but when controlling for the buyer's CF feedback, correlations with DSR are very low, or even negative.

⁴⁶ The effects are statistically highly significant, yet, unfortunately, eBay does not allow us to document this data here, because it could theoretically be related to eBay's profits.

regularities and questions that any theory of the phenomenon will want to address. Going beyond the simple illustration that we present in Appendix B in the context of the experiment, a model could usefully describe, for instance, the noise in feedback, such that more compressed feedback makes it harder to correctly predict quality from the reputation score. Moreover, it would also be useful to endogenize the degree of reciprocity in feedback giving in different institutional environments. This may involve utilizing models of reciprocity, social comparison and group identity (see, Chen et al., 2010, and Chen and Li, 2009, for related observations). Combining theory and empirical studies will further improve our understanding of the role of behavior and design in reputation building.

Our study also has implications for managing the redesign of market trust systems. First, a major challenge in solving marketplace trust problems has to do with possible adverse side-effects or disruptions in path dependencies in migrating to a new system. For example, a redesign of a trust system need respect the fact that reciprocity has positive as well as negative consequences for the feedback system. The giving of feedback is largely a public good, and our data suggest that reciprocity is important for getting mutually satisfactory trades recorded. It is therefore desirable that, in mitigating retaliatory feedback, we strive for a targeted approach rather than one that attempts to remove all forms of reciprocity. Also, by nature, reputation mechanisms are embedded in repeated games, connecting past with future behavior. It was important to the present redesign to maintain certain aspects of the old system, such as the 3-point (conventional) scoring, so that the information collected prior to the change in the system would still be useful in evaluating traders after the changeover, without causing undue confusion.

Second, our laboratory study shows that reciprocal feedback behavior can be channeled, and in a targeted way. The way feedback information is navigated through the system affects whether and how reciprocity influences the candor of feedback. The data show that, compared to a simple open system, both blindness in conventional feedback giving and one-sidedness in a detailed seller rating system increase the information contained in the feedback presented to buyers. As a result, the redesigns likely yield more trust and efficiency in the market, at least in the short-run period that we studied. Additional studies, particularly of longer term effects, should yield further insights.

A third implication has to do with the strength of approaching the problem using complementary methods of analysis. It is the combination and complementary nature of the lab and field data that allows us to be confident in our judgment of likely consequences of institutional changes.⁴⁷ With only field data, it

⁴⁷ As an analogy, consider the work of Humphry Davy, a British chemist of the early 19th century, who summarized the scientific method he followed in the phrase “observation-experiment-analogy.” In a famous series of studies, Humphry went into coal mines to observe the “flammable air” responsible for lethal explosions. He captured some of the gas and took it into his lab for more intensive study. He then drew analogies between field and lab observations. From this he recommended a solution to the coal mines, the famous Davy safety lamp, but also derived a good deal of valuable scientific data about the basic chemistry of methane gas (see Davy 1818). If one substitutes “feedback system” for

would be difficult to unambiguously establish causalities, because both cross- and within platform comparisons involve confounding environment factors. At the same time, laboratory experiments alone do not capture various complexities of the corresponding field environments. They do, however, demonstrate (beyond their benefits as a test-bed for competing designs) that the interaction of institutions and reciprocity is sufficient to cause the robust empirical patterns observed in the various data sets.

eBay introduced ‘detailed seller ratings’ in March and May 2007. Relative to the conventional feedback on eBay, this feedback is more detailed, one-sided and anonymous. The change did not much affect conventional feedback giving, but many traders use the new system to avoid retaliation. This contributes to more reputation dispersion, which in turn leads to improved informativeness. Future studies determining the extent to which the individual components of feedback ratings (detailed, one-sided and anonymous) contribute to the informativeness of future trader performance are a matter of some importance for the efficient application to other Internet and offline market feedback systems.

Naturally, market platforms like eBay continuously monitor and improve trust and trustworthiness on their platform. Motivated by the positive effects of detailed seller ratings, eBay moved ahead and introduced further changes in spring 2008. The most important feature of this more recent change is that sellers are not allowed to submit negative or neutral feedback anymore, only positive. Basically, this is a move to a one-sided feedback system, as found on many business-to-consumer platforms, but still allows for positive reciprocity. Further research will be devoted to how this new change affects the content, timing, and informativeness of feedback. For example, one might expect that, contrary to their behavior in the previous design, more sellers will move first in feedback giving in order to trigger positive reciprocity.

References

- Akerlof, G. (1970), The market for lemons: Quality uncertainty and the market mechanism, *Quarterly Journal of Economics* 84, 488-500.
- Ariely, D., A. Ockenfels, and A. E. Roth (2005), An Experimental Analysis of Ending Rules in Internet Auctions, *The RAND Journal of Economics* 36, 790-809.
- Ba, S. and 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.
- Bajari, P. and A. Hortacısu (2003), The Winner’s Curse, Reserve prices and Endogenous Entry: Empirical Insights from eBay Auctions, *Rand Journal of Economics* 34 (2), 329-355.
- Bajari, P. and A. Hortacısu (2004), Economic Insights from Internet Auctions, *Journal of Economic Literature* 42 (2), 457-486.
- Bar-Isaac, H. and S. Tadelis (2008), Seller Reputation, *Foundations and Trends in Microeconomics* 4, 273-351.
- Bolton, G., B. Greiner, and A. Ockenfels (2010), The Effectiveness of Asymmetric Feedback Systems on eBay. Work in progress.

“methane gas”, “eBay market” for “coal mine”, one has good description of the technique we employ here. See Roth (2002) for another analogy between the economic design of markets and the engineering of suspension bridges.

- Bolton, G., E. Katok, and A. Ockenfels (2004), How Effective are Online Reputation Mechanisms? An Experimental Study, *Management Science* 50 (11), 1587-1602.
- Bolton, G., E. Katok, and A. Ockenfels (2005), Cooperation among Strangers with Limited Information about Reputation, *Journal of Public Economics* 89, 1457-1468.
- Bolton, G. and A. Ockenfels (2008), Does Laboratory Trading Mirror Behavior in Real World Markets? Fair Bargaining and Competitive Bidding on Ebay. Working paper, University of Cologne.
- Bolton, G. and A. Ockenfels (2009), The Limits of Trust in Economic Transactions: Investigations of Perfect Reputation Systems, in: K. S. Cook, C. Snijders, V. Buskens, C. Cheshire (Eds.) *eTrust: Forming Relationships in the Online World*, Russell Sage, 15-36.
- Bolton, G.E., A. Ockenfels, and F. Ebeling (2011), Information Value and Externalities in Reputation Building, *International Journal of Industrial Organization* 29(1), 23-33.
- Brandts, J. and N. Figueras (2003), An Exploration of Reputation Formation in Experimental Games. *Journal of Economic Behavior and Organization* 50, 89-115.
- Bretz, R. D., G. T. Milkovich, and W. Read (1992). The Current State of Performance Appraisal Research and Practice: Concerns, Directions, and Implications. *Journal of Management* 18(2): 321-352.
- Brunner, C., J. K. Goeree, C. A. Holt, and J. O. Ledyard (forthcoming), An Experimental Test of Combinatorial FCC Spectrum Auctions, *American Economic Journal: Microeconomics*.
- Cabral, L. and A. Hortaçsu (2010), The Dynamics of Seller Reputation: Evidence from eBay, *Journal of Industrial Economics* 58, 54-78.
- Camerer, C. F. (2003), *Behavioral Game Theory*, Princeton: Princeton University Press.
- Camerer, C. F. and K. Weigelt (1988), Experimental Tests of a Sequential Equilibrium Reputation Model, *Econometrica* 56, 1-36.
- Chen, K. (2005), An Economics Wind Tunnel: The Science of Business Engineering, in John Morgan (ed.), *Experimental and Behavioral Economics - Advances in Applied Microeconomics*, Volume 13, Elsevier Press, 2005.
- Chen, Y., M. Harper, J. Konstan and S.X. Li (2010), Social Comparisons and Contributions to Online Communities: A Field Experiment on MovieLens. *American Economic Review* 100, 1358-1398.
- Chen, Y., and S.X. Li (2009), Group Identity and Social Preferences. *American Economic Review*, 99(1) 431-457.
- Chen, Y. and T. Sönmez (2006), School choice: An experimental study, *Journal of Economic Theory* 127, 202-231.
- Chwelos, P. and T. Dhar (2007), Differences in "Truthiness" across Online Reputation Mechanisms, Working Paper, Sauder School of Business.
- Cooper, D. and J. Kagel (forthcoming), Other-regarding preferences, in: J. Kagel and A. Roth (eds.), *The Handbook of Experimental Economics*, Volume 2, in preparation.
- Davy, Humphry (1818), *The safety lamp; with researches on flame*, H. Bryer: London (digital version on Google Books).
- Dellarocas, C. (2004), Building Trust On-Line: The Design of Robust Reputation Mechanisms for Online Trading Communities, in: G. Doukidis, N. Mylonopoulos, N. Pouloudi (eds.), *Social and Economic Transformation in the Digital Era*, Idea Group Publishing, Hershey, PA.
- Dellarocas, C. and C. A. Wood (2008), The Sound of Silence in Online Feedback: Estimating Trading Risks in the Presence of Reporting Bias, *Management Science* 54(3), 460-476.
- Dewan, S. and V. Hsu (2001), Trust in Electronic Markets: Price Discovery in Generalist Versus Specialty Online Auctions, mimeo.
- Dulleck, U., R. Kerschbamer, and M. Sutter (forthcoming), The Economics of Credence Goods: On the Role of Liability, Verifiability, Reputation and Competition. *American Economic Review*.
- Eaton, D. H. (2007), The Impact of Reputation Timing and Source on Auction Outcomes, *The B.E. Journal of Economic Analysis & Policy* 7(1), Article 33.
- Ederington, L. H. and M. Dewally (2006), Reputation, Certification, Warranties, and Information as Remedies for Seller-Buyer Information Asymmetries: Lessons from the Online Comic Book Market, *Journal of Business* 79, 693-729.
- Ely, J.C., and J. Valimaki (2003), Bad Reputation, *Quarterly Journal of Economics*, 118, 785-814.
- Fehr, E. and S. Gächter (2000), Fairness and Retaliation: The Economics of Reciprocity, *Journal of Economic Perspectives* 14(3), 159-181.

- Greif, A. (1989), Reputation and Coalitions in Medieval Trade: Evidence on the Maghribi Traders, *Journal of Economic History* 49(4), 857-882.
- Greiner, B. (2004), An Online Recruitment System for Economic Experiments, in: Kurt Kremer, Volker Macho (eds.): *Forschung und wissenschaftliches Rechnen 2003. GWDG Bericht 63*, Göttingen: Ges. für Wiss. Datenverarbeitung, 79-93.
- Greiner, Ben, Axel Ockenfels, and Karim Sadrieh (forthcoming). Internet Auctions. Oxford Handbook of the Digital Economy.
- Grether, D. M., R. M. Isaac, and C. R. Plott (1981), The Allocation of Landing Rights by Unanimity among Competitors, *American Economic Review* 71(2), 166-171.
- Grosskopf, Brit and Rajiv Sarin (2010), Is Reputation Good or Bad? An Experiment, *American Economic Review* 100(5), 2187-2204.
- Güth, W., F. Mengel, and A. Ockenfels (2007), An Evolutionary Analysis of Buyer Insurance and Seller Reputation in Online Markets, *Theory and Decision* 63, 265-282.
- Herrmann, B., C. Thöni and S. Gächter (2008), Antisocial Punishment Across Societies, *Science* 319, 1362-1367.
- Hortaçsu, Ali, F. Asís Martínez-Jerez, and Jason Douglas (2009), The Geography of Trade in Online Transactions: Evidence from eBay and MercadoLibre, *American Economic Journal: Microeconomics* 1(1), 53-74.
- Houser, D. and J. Wooders (2005), Reputation in Auctions: Theory and Evidence from eBay, *Journal of Economics and Management Strategy* 15 (2), 353-369.
- Jian, L., J. MacKie-Mason, and P. Resnick (2010), I Scratched Yours: The Prevalence of Reciprocation in Feedback Provision on eBay, *The B.E. Journal of Economic Analysis & Policy* 10(1), Art. 92.
- Jin, G.Z. and A. Kato (2006). "Price, Quality and Reputation: Evidence from an Online Field Experiment." *RAND Journal of Economics* 37(4), 983-1005.
- Kagel, J. H. and A. E. Roth (2000), The Dynamics of Reorganization in Matching Markets: A Laboratory Experiment Motivated by a Natural Experiment, *Quarterly Journal of Economics* 115(1), 201-235
- Kalyanam, K. and S. McIntyre (2001), Returns to Reputation in Online Auction markets, Retail Workbench Working Paper W-RW01-02, Santa Clara University, Santa Clara, CA.
- Kennes, J. R. and A. Schiff (2007), Simple Reputation Systems, *Scandinavian Journal of Economics* 109, 77-91.
- Klein, T. J., C. Lambert, G. Spagnolo, and K. O. Stahl (2007), Reputation Building in Anonymous Markets: Evidence from eBay, Working Paper, University of Mannheim.
- Klein, B. and K. B. Leffler (1981), The Role of Market Forces in Ensuring Contractual Performance, *Journal of Political Economy* 89, 615-641.
- Kreps, D. M. and R. Wilson (1982), Reputation and Imperfect Information, *Journal of Economic Theory* 27, 253-279.
- Kwasnica, A. M., J. O. Ledyard, D. Porter, and C. DeMartini (2005), A New and Improved Design for Multiobject Iterative Auctions, *Management Science* 51(3), 419-434.
- Lewis, G. (forthcoming), Asymmetric Information, Adverse Selection and Online Disclosure: The Case of eBay Motors, *American Economic Review*.
- Livingston, J. A. (2005), How Valuable is a Good Reputation? A Sample Selection Model of Internet Auctions, *Review of Economics and Statistics* 87 (3), 453-465.
- Livingston, J. A. and W. N. Evans (2004), Do Bidders in Internet Auctions Trust Sellers? A Structural Model of Bidder Behavior on eBay, Working Paper, Bentley College.
- Lucking-Reiley, D., D. Bryan, N. Prasad, and D. Reeves (2007), Pennies from eBay: the Determinants of Price in Online Auctions, *Journal of Industrial Economics* 55(2), 223-233.
- McDonald, C. G. and V. C. Slawson, Jr. (2002), Reputation in an Internet Auction Market, *Economic Inquiry* 40 (3), 633-650.
- Melnik, M. I. and J. Alm (2002), Does a Seller's Reputation Matter? Evidence from eBay Auctions, *Journal of Industrial Economics* 50 (3), 337-349.
- Milgrom, P. (2004), *Putting Auction Theory to Work*, Cambridge University Press.
- Milgrom, P., D. North, and B. Weingast (1990), The Role of Institutions in the Revival of Trade: The Medieval Law Merchant, *Economics and Politics* 2, 1-23.

- Muniz, J., E. Garcia-Cueto, and L. M. Lozano (2005), Item formation and the psychometric properties of the Eysenck Personality Questionnaire, *Personality and Individual Differences* 38, 61-69.
- Neral, J. and J. Ochs (1992), The Sequential Equilibrium Theory of Reputation Building: A Further Test, *Econometrica* 60(5), 1151-1169.
- Niederle, M. and A. E. Roth (2005), The Gastroenterology Fellowship Market: Should there be a Match?, *American Economic Review Papers & Proceedings* 95 (2), 372-375.
- Niederle, M. and A. E. Roth (2009), Market Culture: How Rules Governing Exploding Offers Affect Market Performance, *American Economic Journal: Microeconomics* 1 (2), 199-219.
- Nowak, M.A. and K. Sigmund (1998), Evolution of Indirect Reciprocity by Image Scoring. *Nature*, 393, 573-577.
- Nunnally, J. C. (1978), Psychometric theory (2nd ed.), McGraw-Hill: New York.
- Ockenfels, A. (2003), Reputationsmechanismen auf Internet-Marktplattformen: Theorie und Empirie, *Zeitschrift für Betriebswirtschaft* 73 (3), 295-315.
- Ockenfels, A., and P. Resnick (2010), Negotiating Reputations, Working Paper.
- Ockenfels, A. and A. E. Roth (2006), Late and Late and multiple bidding in second price Internet auctions: Theory and evidence concerning different rules for ending an auction, *Games and Economic Behavior* 55, 297-320.
- Oppenheim, A.N. (2000), Questionnaire Design, Interviewing and Attitude Measurement, Continuum: London and New York.
- Prendergast, C. (1999). The Provision of Incentives in Firms. *Journal of Economic Literature* 37(1): 7-63.
- Prendergast, C. and R. H. Topel (1993). Discretion and Bias in Performance Evaluation. *European Economic Review* 37(2-3): 355-365.
- Reichling, F. (2004), Effects of Reputation Mechanisms on Fraud Prevention in eBay Auctions, Working Paper, Stanford University.
- Resnick, P. and R. Zeckhauser (2002), Trust Among Strangers in Internet Transactions: Empirical Analysis of eBay's Reputation System, in: Michael R. Baye (ed.), *The Economics of the Internet and E-Commerce* (Advances in Applied Microeconomics, Vol. 11), JAI Press.
- Resnick, P., R. Zeckhauser, E. Friedman, and K. Kuwabara (2000), Reputation Systems, *Communications of the ACM* 43 (12), 45-48.
- Resnick, P., R. Zeckhauser, J. Swanson, and L. Lockwood (2006), The Value of Reputation on eBay: A Controlled Experiment, *Experimental Economics* 9, 79-101.
- Roth, A. E. (2002), The Economist as Engineer: Game Theory, Experimentation, and Computation as Tools for Design Economics, Fisher-Schultz Lecture, *Econometrica* 70(4), 1341-1378.
- Roth, A. E. (2008), What have we learned from market design?, Hahn Lecture, *Economic Journal* 118, 285-310.
- Roth, A. E. and A. Ockenfels (2002), Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet, *American Economic Review* 92 (4), 1093-1103.
- Selten, R. and R. Stöcker (1986), End behavior in sequences of finite prisoner's dilemma supergames, *Journal of Economic Behavior and Organization* 7, 47-70.
- Sutter, M., S. Haigner and M. Kocher (2010). Choosing the Stick or the Carrot? – Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies* 77, 1540-1566.
- Wilson, R. (1985), Reputations in Games and Markets, in: A. E. Roth (ed.), *Game-Theoretic Models of Bargaining*, Cambridge University Press, Cambridge, UK, 27-62.